

ϵ -Subgradient Algorithms for Bilevel Convex Optimization

Elias S. Helou and Lucas E. A. Simões

the date of receipt and acceptance should be inserted later

Abstract This paper introduces and studies the convergence properties of a new class of explicit ϵ -subgradient methods for the task of minimizing a convex function over the set of minimizers of another convex minimization problem. The general algorithm specializes to some important cases, such as first-order methods applied to a varying objective function, which have computationally cheap iterations.

We present numerical experimentation regarding certain applications where the theoretical framework encompasses efficient algorithmic techniques, enabling the use of the resulting methods to solve very large practical problems arising in tomographic image reconstruction.

Mathematics Subject Classification (2000) 65K10, 90C25, 90C56

Keywords convex optimization, nondifferentiable optimization, bilevel optimization, ϵ -subgradients

1 Introduction

Our aim in the present paper is to solve a *bilevel* or *hierarchical* optimization problem of the form

$$\begin{aligned} \min \quad & f_1(\mathbf{x}) \\ \text{s. t.} \quad & \mathbf{x} \in \operatorname{argmin}_{\mathbf{y} \in X_0} f_0(\mathbf{y}), \end{aligned} \tag{1}$$

E.S. Helou was supported by FAPESP grants 2013/07375-0 and 2013/16508-3 and CNPq grant 311476/2014-7. L.E.A. Simões was supported by FAPESP grant 2013/14615-7.

Elias S. Helou

Department of Applied Mathematics and Statistics, Institute of Mathematical Sciences and Computation – USP, São Carlos, elias@icmc.usp.br

Lucas E. A. Simões

Institute of Mathematics, Statistics and Scientific Computing – UNICAMP, Campinas

where $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ ($i \in \{0, 1\}$) are convex functions and X_0 is a nonempty closed convex set.

Bilevel problems like (1) have already been considered in the literature. For example, for the case $X_0 = \mathbb{R}^n$, Cabot [8] suggests the use of the following algorithm:

$$-\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\lambda_k} \in \partial_{\epsilon_k}(f_0 + \eta_k f_1)(\mathbf{x}_{k+1}), \quad (2)$$

where $\partial_{\epsilon} f(\mathbf{x})$ is the ϵ -subdifferential of f at \mathbf{x} :

$$\partial_{\epsilon} f(\mathbf{x}) := \{\mathbf{v} : f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{v}^T(\mathbf{y} - \mathbf{x}) - \epsilon, \quad \forall \mathbf{y} \in \mathbb{R}^n\},$$

$\eta_k \rightarrow 0^+$ and λ_k is a nonnegative stepsize. Such iterations are reminiscent of approximate proximal methods, in the sense that (2) is equivalent to (we follow the notation of [8]):

$$\mathbf{x}_{k+1} \in \epsilon_k\text{-argmin}_{\mathbf{x} \in X_0} \left\{ \frac{1}{2\lambda_k} \|\mathbf{x} - \mathbf{x}_k\|^2 + f_0(\mathbf{x}) + \eta_k f_1(\mathbf{x}) \right\}. \quad (3)$$

While method (2) is powerful and conceptually simple, its application may be complicated by the implicit formulation. Assuming differentiability, iteration (2) can also be interpreted as a discretization of the continuous dynamical system

$$\dot{\mathbf{x}}(t) + \nabla f_0(\mathbf{x}(t)) + \eta(t) \nabla f_1(\mathbf{x}(t)) = 0.$$

Another way of discretizing this system is to do it explicitly, that is, to use iterations similar to

$$-\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\lambda_k} \in \partial_{\epsilon_k}(f_0 + \eta_k f_1)(\mathbf{x}_k). \quad (4)$$

or in two steps:

$$\begin{aligned} -\frac{\mathbf{x}_{k+1/2} - \mathbf{x}_k}{\lambda_k} &\in \partial_{\epsilon_k^0} f_0(\mathbf{x}_k) \\ -\frac{\mathbf{x}_{k+1} - \mathbf{x}_{k+1/2}}{\lambda_k \eta_k} &\in \partial_{\epsilon_k^1} f_1(\mathbf{x}_{k+1/2}). \end{aligned} \quad (5)$$

Among the consequences of the results we will present in this paper, there are sufficient conditions on the sequences $\{\lambda_k\}$, $\{\epsilon_k\}$ and $\{\eta_k\}$ for the convergence of iterations (5) to the solution of problem (1). In fact, convergence of iterations (4) could also be proven using our abstract results, but would require more restrictive conditions on ϵ_k and we will keep this topic off the present paper.

Despite the fact that algorithms (2) and (5) are formally very similar, they differ significantly in both practical and theoretical aspects. At the practical side, implementation of an explicit iteration like (5) requires little more than evaluation of suitable ϵ -subgradients. On the other hand, while strong convexity makes this minimization perhaps be more computationally amenable than approximately optimizing $f_0 + \eta_k f_1$, computing (3) is still a nontrivial task. Furthermore, even if it allows for some tolerance ϵ_k in the optimization

subproblem, current theory requires $\sum_{k=0}^{\infty} \epsilon_k < \infty$ to ensure convergence of (2), which means that this tolerance will decrease quickly.

Arguably, an algorithm such as (5) would be among the easiest to implement methodology for solving (1) in the general case. For example, Solodov [34, 35] has also provided algorithms for such bilevel problems: in [34], where only the differentiable case is considered, the proposed algorithm has the form (4) with $\epsilon_k \equiv 0$ and λ_k selected through a line search with a sufficient decrease criterion based on the value of $f_0 + \eta_k f_1$. While this procedure is still simple to implement on the differentiable case, descent directions are harder to be found in the presence of nondifferentiability and in [35] a bundle technique is used to this end. This approach requires sophisticated quadratic solvers, which complicates implementations. Another, partial and perturbed, descent method was developed by Helou and De Pierro [22], where only sufficient decrease of f_0 , assumed to be smooth, is enforced, alleviating the need of a descent direction for f_1 . But this technique may still require multiple evaluations of f_0 , while an iterative process like (5), differently, does not require any f_i value because it does not rely on descent criteria. Furthermore, it should be remarked that the fact that (5) allows for inexactness in the computation has positive impact in algorithmic performance, which we will illustrate through experimental work. An approach currently available in the literature which is similar to a special case of the techniques that we can analyze within our framework can be found in [7]. In this work the results appear to be restricted to monotone methods for a quadratic residual function, but a practical stopping criterion based on the discrepancy principle is given. Another recent work dealing with this kind of problem is [1] where a first order algorithm is proposed and convergence analysis including rates is provided. While we do not provide convergence rates, the theory we present requires less hypothesis on the objective functions and seems to give rise to a wider range of practical algorithms.

1.1 Contributions and Outline of the Paper

The main contribution of the present paper is as follows. The thread lead by Cabot and followed by Solodov is based on the idea of applying classical convex minimization algorithms to the ever-changing objective function $f_0 + \eta_k f_1$. It started with the “tight” near-minimization from [8] and evolved to the less stringent sufficient decrease policy of [34, 35]. We here pave this way one step further by showing that the same principle is applicable to the more anarchic nonmonotone ϵ -subgradient techniques.

Several unconstrained optimization algorithms have iterations that can be described as ϵ -subgradient steps, among which we can mention the incremental subgradient methods [29], the aggregated incremental gradient of [6] when applied to the nondifferentiable case [22], the recent incremental proximal method [3], and Polyak’s heavy ball method (let $\tilde{\nabla}f(\mathbf{x}) \in \partial f(\mathbf{x})$):

$$\mathbf{x}_{k+1} := \mathbf{x}_k - \lambda_k (\tilde{\nabla}f(\mathbf{x}_k) + \alpha(\mathbf{x}_k - \mathbf{x}_{k-1})).$$

The theory we develop will, therefore, cope with all the just mentioned cases simultaneously. That is, we show that application of any of these algorithms to the varying objective function $f_0 + \eta_k f_1$ will converge to the solution of the bilevel problem (1) under assumptions on the stepsize λ_k which are not much different from those required in the one level case.

We present numerical experimentation showing the effectiveness of the technique when applied to high-resolution micro-tomographic image reconstruction from simulated and from real synchrotron radiation illumination projection data. In this case the amount of data and the number of variables is very large, requiring efficient algorithms with computationally cheap iterations. In this context, important practical contributions are the introduction of certain perturbed, FISTA-inspired [2] algorithms, resulting in effective methods for problems like (1) with Lipschitz-differentiable f_0 . Furthermore, when solving an instance with a non-differentiable f_0 composed as a sum of many convex functions, incremental techniques are very efficient in the first iterations, also resulting in good algorithmic performance. Both the theoretical analysis and the application of these practical algorithms to the bilevel problem (1) are new.

Our methods can be seen as perturbations of classical algorithms, in the spirit of the superiorization approach [19]. However, we show more powerful convergence results because we impose some structure on what would otherwise be called a superiorization sequence. We believe that this is a major contribution of the paper because opens the possibility of pursuing bilevel results alongside with superiorization techniques.

2 Theoretical Analysis

2.1 Stepsize Considerations

Recall from the theory of ϵ -subgradient methods [12] for the one level case (that is, problem (1) with $f_1 \equiv 0$) that convergence of iterations (4) to a solution, under mild extra assumptions of subgradient boundedness, can be ensured with slowly diminishing stepsizes satisfying:

$$\sum_{k=0}^{\infty} \lambda_k = \infty \quad \text{and} \quad \lambda_k \rightarrow 0^+.$$

The non-summability hypothesis seems necessary. Vanishing stepsizes, however, may have the negative effect of slowing down asymptotic convergence of the algorithm. Therefore, owing to its computationally cheap iteration, methods like (4) are usually thought to be most competitive when the problem size is very large, or when highly accurate solutions are not required. However, in some important particular (with smooth primary objective function f_0) cases supported by the theory developed here, the stepsize λ_k does not necessarily have to vanish and in such applications we obtain reasonably fast algorithms. Both stepsize regimes (decreasing and non-decreasing) are evaluated in the

experimental work we present and we shall see that incremental techniques are efficient too in the cases where its good characteristics apply, even if with vanishing stepsizes.

For the classical (one level) convex optimization problem, proximal methods [30] require $\lambda_k \geq \underline{\lambda}$ for some $\underline{\lambda} > 0$. For the bilevel case, the same stepsize requirement, with an extra upper boundedness assumption, i.e.,

$$0 < \underline{\lambda} \leq \lambda_k \leq \bar{\lambda}, \quad (6)$$

would ensure convergence of (2) to the optimizer of (1). For such results to hold [8], the extra assumption of a *slow control*:

$$\sum_{k=0}^{\infty} \eta_k = \infty \quad \text{and} \quad \eta_k \rightarrow 0^+, \quad (7)$$

was made in order to ensure that the influence of f_1 throughout the iterations was strong enough while still becoming arbitrarily small. The alternative form below is more appropriate to us this time, as it will generalize immediately to our algorithms:

$$\sum_{k=0}^{\infty} \lambda_k \eta_k = \infty \quad \text{and} \quad \eta_k \rightarrow 0^+. \quad (8)$$

Notice that if (6) holds, then (8) is equivalent to (7). However, because the net contribution to each iteration from $\partial f_1(\mathbf{x}_k)$ in algorithm (4) is actually $O(\lambda_k \eta_k)$, (8) generalizes to the case $\lambda_k \rightarrow 0^+$, while (7) does not.

2.2 Formal Algorithm Description

The stepping stone of our analysis will be an abstract three-step algorithm given as follows:

$$\begin{aligned} \mathbf{x}_{k+1/3} &:= \mathcal{O}_{f_0}(\lambda_k, \mathbf{x}_k); \\ \mathbf{x}_{k+2/3} &:= \mathcal{O}_{f_1}(\mu_k, \mathbf{x}_{k+1/3}); \\ \mathbf{x}_{k+1} &:= \mathcal{P}_{X_0}(\mathbf{x}_{k+2/3}), \end{aligned} \quad (9)$$

where the operators \mathcal{O}_{f_i} for $i \in \{0, 1\}$ have specific conceptual roles and must satisfy certain corresponding properties, which we will discuss right next, and, for a nonempty convex and closed set X , \mathcal{P}_X is the projector:

$$\mathcal{P}_X(\mathbf{x}) := \operatorname{argmin}_{\mathbf{y} \in X} \|\mathbf{x} - \mathbf{y}\|.$$

Sequences $\{\lambda_k\}$ and $\{\mu_k\}$ are stepsize sequences. The first one plays, in this abstract setting, the same role it plays in iterative scheme (4), while sequence $\{\mu_k\}$ should be identified with $\{\lambda_k \eta_k\}$. Therefore, application of $\mu_k = \lambda_k \eta_k$ in (8) leads immediately to

$$\sum_{k=0}^{\infty} \mu_k = \infty \quad \text{and} \quad \frac{\mu_k}{\lambda_k} \rightarrow 0^+.$$

Let us then describe the properties required for the *optimality operators* \mathcal{O}_{f_i} . The imposed characteristics are easy to meet, as we later illustrate.

Property 1 There is $\beta > 0$ such that for any $\lambda \geq 0$ and for all $\mathbf{x}_{k+i/3}, \mathbf{y} \in \mathbb{R}^n$, and $i \in \{0, 1\}$:

$$\|\mathcal{O}_{f_i}(\lambda, \mathbf{x}_{k+i/3}) - \mathbf{y}\|^2 \leq \|\mathbf{x}_{k+i/3} - \mathbf{y}\|^2 - \beta\lambda(f_i(\mathcal{O}_{f_i}(\lambda, \mathbf{x}_{k+i/3})) - f_i(\mathbf{y})) + \lambda\rho_i(\lambda, k),$$

where $\rho_i(\lambda, k)$ represents an error term, with properties to be describe later.

Below the description of the next property, we give an example of a class of operators which satisfy this condition. Furthermore, Subsections 2.4, 2.5, and 3.4 bring four other instances that will be used in the experimental part of the paper: the projected gradient, the incremental subgradient, the proximal map and the iterated subgradient step. In fact, the key utility of this abstract definition is to be able to encompass several useful classical optimization steps while still ensuring sufficient qualities in order to provide convergence results. For this to be true, the error term will have to be controlled in a specific way, but, for every case we have found, the error term magnitude is bounded by a constant times the stepsize and this way we can always obtain convergent algorithms by selecting proper stepsize sequences.

Property 2 There exists $\gamma > 0$ such that

$$\|\mathbf{x} - \mathcal{O}_{f_i}(\lambda, \mathbf{x})\|_2 \leq \lambda\gamma.$$

Property 1 guarantees that, going from some fixed \mathbf{x} , the operator \mathcal{O}_{f_i} will approach a point \mathbf{y} with improved f_i value if only the result of the operation does not have a better f_i value than \mathbf{y} and the stepsize λ is small enough. Property 2 is no more than a boundedness assumption on the operators which makes sure that the stepsize controls the magnitude of the movement.

These can be derived from somewhat standard hypothesis for ϵ -subgradient algorithms (see, e.g., [12]) and, as such, a plethora of concrete realizations of such operators \mathcal{O}_f is possible, the most obvious being ϵ -subgradient steps:

$$\mathcal{S}_f(\lambda, \mathbf{x}) := \mathbf{x} - \lambda\tilde{\nabla}_\epsilon f(\mathbf{x}),$$

where $\tilde{\nabla}_\epsilon f(\mathbf{x}) \in \partial_\epsilon f(\mathbf{x})$. In this case we have:

$$\|\mathcal{S}_f(\lambda, \mathbf{x}) - \mathbf{y}\|_2^2 \leq \|\mathbf{x} - \mathbf{y}\|_2^2 - 2\lambda(f(\mathbf{x}) - f(\mathbf{y})) + \lambda(\lambda\|\tilde{\nabla}_\epsilon f(\mathbf{x})\|_2^2 + 2\epsilon). \quad (10)$$

Denote $\mathbf{z} = \mathbf{x} - \lambda\tilde{\nabla}_\epsilon f(\mathbf{x})$. Then, convexity leads to

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \tilde{\nabla}f(\mathbf{z})^T(\mathbf{x} - \mathbf{z}) = f(\mathbf{z}) + \lambda\tilde{\nabla}f(\mathbf{z})^T\tilde{\nabla}_\epsilon f(\mathbf{x}),$$

where $\tilde{\nabla}f(\mathbf{z}) \in \partial f(\mathbf{z})$. Multiplying the above inequality by -2λ we get

$$-2\lambda f(\mathbf{x}) \leq -2\lambda\left(f(\mathcal{S}_f(\lambda, \mathbf{x})) + \lambda\tilde{\nabla}f(\mathcal{S}_f(\lambda, \mathbf{x}))^T\tilde{\nabla}_\epsilon f(\mathbf{x})\right)$$

This with (10) gives

$$\begin{aligned} \|\mathcal{S}_f(\lambda, \mathbf{x}) - \mathbf{y}\|_2^2 &\leq \|\mathbf{x} - \mathbf{y}\|_2^2 - 2\lambda(f(\mathcal{S}_f(\lambda, \mathbf{x})) - f(\mathbf{y})) \\ &\quad + \lambda(\lambda\|\tilde{\nabla}_\epsilon f(\mathbf{x})\|_2^2 + 2\epsilon - 2\lambda\tilde{\nabla} f(\mathcal{S}_f(\lambda, \mathbf{x}))^T \tilde{\nabla}_\epsilon f(\mathbf{x})), \end{aligned} \quad (11)$$

so that we can satisfy Properties 1 and 2 for $\mathcal{O}_f = \mathcal{S}_f$ if we further assume ϵ -subgradient boundedness (and consequently subgradient boundedness).

A straightforward generalization of the argument leading from (10) to (11), omitted for brevity, results in the following statement, which will be useful later:

Proposition 1 *Assume an operator $\mathcal{O}_f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfies, for $\lambda > 0$*

$$\|\mathcal{O}_f(\lambda, \mathbf{x}) - \mathbf{y}\|_2^2 \leq \|\mathbf{x} - \mathbf{y}\|_2^2 - 2\lambda(f(\mathbf{x}) - f(\mathbf{y})) + \lambda\varrho(\lambda),$$

where $\varrho(\lambda)$ is an error term, and

$$\|\mathcal{O}_f(\lambda, \mathbf{x}) - \mathbf{x}\| \leq \lambda\gamma,$$

for some $\gamma > 0$. Then, we have:

$$\begin{aligned} \|\mathcal{O}_f(\lambda, \mathbf{x}) - \mathbf{y}\|_2^2 &\leq \|\mathbf{x} - \mathbf{y}\|_2^2 - 2\lambda(f(\mathcal{O}_f(\lambda, \mathbf{x})) - f(\mathbf{y})) \\ &\quad + \lambda(\varrho(\lambda) + 2\lambda\gamma\|\tilde{\nabla} f(\mathcal{O}_f(\lambda, \mathbf{x}))\|), \end{aligned}$$

where $\tilde{\nabla} f(\mathcal{O}_f(\lambda, \mathbf{x})) \in \partial f(\mathcal{O}_f(\lambda, \mathbf{x}))$.

Our analysis will focus on algorithms more general than (4), allowing simple constraint sets X_0 to be handled. We recall that the projection onto a nonempty convex closed set X_0 satisfies:

Property 3 For all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in X_0$, we have

$$\|\mathcal{P}_{X_0}(\mathbf{x}) - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\|. \quad (12)$$

2.3 Convergence Results

We introduce some simplifying notations:

- f_i^* is the optimal value of f_i over X_i for $i \in \{0, 1\}$ where;
- X_0 is given and $X_{i+1} := \{\mathbf{x} \in X_i : f_i(\mathbf{x}) = f_i^*\}$ for $i \in \{0, 1\}$;
- $[x]_+ := \max\{0, x\}$;
- $d_X(\mathbf{x}) := \|\mathbf{x} - \mathcal{P}_X(\mathbf{x})\|$.

Our first result shows convergence of the iterates to the set of minimizers of f_0 over X_0 . We next prove convergence to the set of minimizers of f_1 over X_1 . Both of these preliminary results contain certain technical and some apparently strong hypothesis. We subsequently weaken and clarify such *ad hoc* requirements in order to obtain our main results.

Proposition 2 Assume that $X_1 \neq \emptyset$, X_1 is bounded (or $\{\mathbf{x}_k\}$ is bounded) $\sum_{i=0}^{\infty} \lambda_k = \infty$, \mathcal{O}_{f_0} and \mathcal{O}_{f_1} satisfy Property 1, \mathcal{O}_{f_1} satisfy also Property 2, $f_1(\mathbf{x}_{k+2/3}) \geq \underline{f} > -\infty$, $\|\mathbf{x}_k - \mathbf{x}_{k+1/3}\| \rightarrow 0$, $\rho_0(\lambda_k, k) \rightarrow 0$, $\rho_1(\mu_k, k) \leq \bar{\rho}_1 < \infty$, $\mu_k \rightarrow 0$, and $\mu_k/\lambda_k \rightarrow 0$. Suppose also that there exists M such that $\forall k \in \mathbb{N}$ there is $\mathbf{v}_k \in \partial f_0(\mathbf{x}_k)$ for which $\|\mathbf{v}_k\| < M$, then we have

$$\lim_{k \rightarrow \infty} d_{X_1}(\mathbf{x}_k) = 0.$$

Proof First notice that $\mu_k \rightarrow 0$ and Property 2 imply $\|\mathbf{x}_{k+1/3} - \mathbf{x}_{k+2/3}\| \rightarrow 0$. Then we take into consideration the non-expansiveness of the projection and of Property 1 of \mathcal{O}_{f_0} and \mathcal{O}_{f_1} , there holds, for $\mathbf{y} \in X_1$:

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 &\leq \|\mathbf{x}_{k+2/3} - \mathbf{y}\|^2 \\ &\leq \|\mathbf{x}_{k+1/3} - \mathbf{y}\|^2 - \beta\mu_k(f_1(\mathbf{x}_{k+2/3}) - f_1(\mathbf{y})) + \mu_k\rho_1(\mu_k, k) \\ &\leq \|\mathbf{x}_k - \mathbf{y}\|^2 - \beta\lambda_k(f_0(\mathbf{x}_{k+1/3}) - f_0^*) + \lambda_k\rho_0(\lambda_k, k) \\ &\quad - \beta\mu_k(f_1(\mathbf{x}_{k+2/3}) - f_1(\mathbf{y})) + \mu_k\rho_1(\mu_k, k). \end{aligned} \quad (13)$$

Denote $N = f_1(\mathbf{y}) - \underline{f}$, thereby simplifying the above expression to:

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 &\leq \|\mathbf{x}_k - \mathbf{y}\|^2 - \beta\lambda_k(f_0(\mathbf{x}_{k+1/3}) - f_0^*) \\ &\quad + \lambda_k\rho_0(\lambda_k, k) + \mu_k(\beta N + \rho_1(\mu_k, k)). \end{aligned} \quad (14)$$

Then, the boundedness of $\partial f_0(\mathbf{x}_k)$ leads to

$$f_0(\mathbf{x}_{k+1/3}) \geq f_0(\mathbf{x}_k) - M\|\mathbf{x}_{k+1/3} - \mathbf{x}_k\|,$$

which together with (14) gives

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 &\leq \|\mathbf{x}_k - \mathbf{y}\|^2 - \beta\lambda_k(f_0(\mathbf{x}_k) - f_0^*) \\ &\quad + \lambda_k(\rho_0(\lambda_k, k) + \beta M\|\mathbf{x}_{k+1/3} - \mathbf{x}_k\|) + \mu_k(\beta N + \rho_1(\mu_k, k)). \end{aligned} \quad (15)$$

We shall denote, for $\delta \geq 0$:

$$X_1^\delta := \{\mathbf{x}_k : f_0(\mathbf{x}_k) \leq f_0^* + \delta\}.$$

Notice that if X_1 is bounded (or $\{\mathbf{x}_k\}$ is bounded), then X_1^δ is bounded. Therefore, the following quantity is well defined:

$$\Delta_1(\delta) := \sup_{\mathbf{x} \in X_1^\delta} d_{X_1}(\mathbf{x}).$$

Furthermore, we have

$$\lim_{\eta \rightarrow 0} \Delta_1(\delta + \eta) = \Delta_1(\delta) \quad \text{and} \quad \Delta_1(0) = 0.$$

Let δ be any positive real number and consider, with $\rho_1(\mu_k, k) \leq \bar{\rho}_1$, $\mu_k/\lambda_k \rightarrow 0$, $\|\mathbf{x}_{k+1/3} - \mathbf{x}_k\| \rightarrow 0$, and $\rho_0(\lambda_k, k) \rightarrow 0$ in mind, that k_0 is large enough such that $k \geq k_0$ implies

$$\rho_0(\lambda_k, k) + \beta M \|\mathbf{x}_{k+1/3} - \mathbf{x}_k\| < \beta \frac{\delta}{3}, \quad \text{and} \quad \frac{\mu_k}{\lambda_k} (\beta N + \rho_1(\mu_k, k)) < \beta \frac{\delta}{3}. \quad (16)$$

Then, two situations may occur:

1. $d_{X_1}(\mathbf{x}_k) \geq \Delta_1(\delta)$;
2. $d_{X_1}(\mathbf{x}_k) < \Delta_1(\delta)$.

Let us first suppose that Case 1 holds, that is $f_0(\mathbf{x}_k) - f_0^* \geq \delta$. Then, for $k \geq k_0$, from (15) and (16) we get:

$$\|\mathbf{x}_{k+1} - \mathbf{y}\|^2 \leq \|\mathbf{x}_k - \mathbf{y}\|^2 - \beta \lambda_k \frac{\delta}{3}.$$

In particular,

$$d_{X_1}(\mathbf{x}_{k+1})^2 \leq \|\mathbf{x}_{k+1} - \mathcal{P}_{X_1}(\mathbf{x}_k)\|^2 \leq d_{X_1}(\mathbf{x}_k)^2 - \beta \lambda_k \frac{\delta}{3}.$$

Therefore, since $\sum_{k=0}^{\infty} \lambda_k = \infty$, there must exist an arbitrarily large $k_1 \geq k_0$ such that $d_{X_1}(\mathbf{x}_k) < \Delta_1(\delta)$.

Now, let us notice that, because of (12)

$$\begin{aligned} d_{X_1}(\mathbf{x}_{k+1}) &\leq \|\mathbf{x}_{k+1} - \mathcal{P}_{X_1}(\mathbf{x}_k)\| \\ &\leq \|\mathbf{x}_{k+2/3} - \mathcal{P}_{X_1}(\mathbf{x}_k)\| \\ &\leq d_{X_1}(\mathbf{x}_k) + \|\mathbf{x}_k - \mathbf{x}_{k+2/3}\|. \end{aligned} \quad (17)$$

Given the hypothesis, we may assume that k_0 is large enough such that, in addition to (16), we have also

$$\|\mathbf{x}_k - \mathbf{x}_{k+2/3}\| \leq \delta.$$

Therefore, for $k > k_1$, there holds:

$$d_{X_1}(\mathbf{x}_k) \leq \Delta_1(\delta) + \delta.$$

Since $\delta > 0$ was arbitrary and $\lim_{\delta \rightarrow 0} \Delta_1(\delta) = 0$, the claim is proven. \square

Proposition 3 Assume $X_2 \neq \emptyset$, X_2 is bounded (or $\{\mathbf{x}_k\}$ is bounded), that $\mu_k \rightarrow 0$, $\sum_{i=0}^{\infty} \mu_k = \infty$, \mathcal{O}_{f_0} and \mathcal{O}_{f_1} satisfy Property 1, \mathcal{O}_{f_1} also satisfies Property 2, $\lambda_k[f_0^* - f_0(\mathbf{x}_{k+1/3})]/\mu_k \rightarrow 0$, $d_{X_1}(\mathbf{x}_k) \rightarrow 0$, $\|\mathbf{x}_k - \mathbf{x}_{k+1/3}\| \rightarrow 0$, $\lambda_k \rho_0(\lambda_k, k)/\mu_k \rightarrow 0$ and $\rho_1(\mu_k, k) \rightarrow 0$. Suppose also that there exists an M such that $\forall k \in \mathbb{N}$ there are $\mathbf{v}_k \in \partial f_0(\mathbf{x}_k)$ and $\mathbf{w}_k \in \partial f_1(\mathcal{P}_{X_0}(\mathbf{x}_k))$ for which $\|\mathbf{v}_k\| < M$ and $\|\mathbf{w}_k\| \leq M$, then we have

$$\lim_{k \rightarrow \infty} d_{X_2}(\mathbf{x}_k) = 0.$$

Proof Notice for later reference that just like in Proposition 2, the hypotheses imply that $\|\mathbf{x}_{k+1/3} - \mathbf{x}_{k+2/3}\| \rightarrow 0$. Now, if we use (13) with $\mathbf{y} \in X_2 \subset X_1$, we get:

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 &\leq \|\mathbf{x}_k - \mathbf{y}\|^2 - \beta\mu_k(f_1(\mathbf{x}_{k+2/3}) - f_1(\mathbf{y})) \\ &\quad + \lambda_k\rho_0(\lambda_k, k) + \mu_k\rho_1(\mu_k, k) + \beta\lambda_k[f_0^* - f(\mathbf{x}_{k+1/3})]_+. \end{aligned} \quad (18)$$

Now let $\tilde{\nabla}f_1(\mathbf{x}_k) \in \partial f_1(\mathbf{x}_k)$ and then notice that convexity of f_1 , Cauchy-Schwarz inequality and the boundedness assumption on $\partial f_1(\mathbf{x}_k)$ lead to:

$$\begin{aligned} f_1(\mathbf{x}_{k+2/3}) &\geq f_1(\mathbf{x}_k) + \tilde{\nabla}f_1(\mathbf{x}_k)^T(\mathbf{x}_{k+2/3} - \mathbf{x}_k) \\ &\geq f_1(\mathbf{x}_k) - \|\tilde{\nabla}f_1(\mathbf{x}_k)\|\|\mathbf{x}_{k+2/3} - \mathbf{x}_k\| \\ &\geq f_1(\mathbf{x}_k) - M\|\mathbf{x}_{k+2/3} - \mathbf{x}_k\|. \end{aligned} \quad (19)$$

Then, using (19) in (18) it is possible to obtain:

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{y}\|^2 &\leq \|\mathbf{x}_k - \mathbf{y}\|^2 - \beta\mu_k(f_1(\mathbf{x}_k) - f_1^*) + \lambda_k\rho_0(\lambda_k, k) \\ &\quad + \mu_k\rho_1(\mu_k, k) + \beta\lambda_k[f_0^* - f_0(\mathbf{x}_{k+1/3})]_+ + \beta\mu_k M\|\mathbf{x}_{k+2/3} - \mathbf{x}_k\|. \end{aligned} \quad (20)$$

Similarly to the Δ_1 notation introduced above, we will denote, for $\delta \geq 0$:

$$X_2^\delta := \{\mathbf{x}_k : f_1(\mathcal{P}_{X_1}(\mathbf{x}_k)) \leq f_1^* + \delta\}.$$

Notice that if X_2 is bounded (or $\{\mathbf{x}_k\}$ is bounded) and $d_{X_1}(\mathbf{x}_k)$ is also bounded, then X_2^δ is bounded. Therefore, the following quantity is well defined:

$$\Delta_2(\delta) := \sup_{\mathbf{x} \in X_2^\delta} d_{X_2}(\mathbf{x}).$$

Furthermore, we have

$$\lim_{\eta \rightarrow 0} \Delta_2(\delta + \eta) = \Delta_2(\delta) \quad \text{and} \quad \Delta_2(0) = 0.$$

Given the hypothesis, for any fixed $\delta > 0$, there is k_0 such that $k \geq k_0$ implies that

$$\begin{aligned} \frac{\lambda_k\rho_0(\lambda_k, k)}{\mu_k} &< \beta\frac{\delta}{5}, \quad \rho_1(\mu_k, k) < \beta\frac{\delta}{5}, \quad \frac{\lambda_k[f_0^* - f(\mathbf{x}_{k+1/3})]_+}{\mu_k} < \frac{\delta}{5}, \\ \text{and} \quad \mu_k M\|\mathbf{x}_{k+2/3} - \mathbf{x}_k\| &< \frac{\delta}{5}. \end{aligned} \quad (21)$$

We from now on assume $k > k_0$ and split in two different possibilities:

1. $f_1(\mathbf{x}_k) > f_1^* + \delta$;
2. $f_1(\mathbf{x}_k) \leq f_1^* + \delta$.

We start by analyzing Case 1. Using (21) in (20) we get, for $\mathbf{y} \in X_2$:

$$\|\mathbf{x}_{k+1} - \mathbf{y}\|^2 < \|\mathbf{x}_k - \mathbf{y}\|^2 - \beta\mu_k \frac{\delta}{5}.$$

In particular:

$$d_{X_2}(\mathbf{x}_{k+1})^2 \leq \|\mathbf{x}_{k+1} - \mathcal{P}_{X_2}(\mathbf{x}_k)\|^2 < d_{X_2}(\mathbf{x}_k)^2 - \beta\mu_k \frac{\delta}{5}.$$

Because of $\sum_{k=0}^{\infty} \mu_k = \infty$, this inequality means that there is an arbitrarily large k_1 such that $f_1(\mathbf{x}_{k_1}) \leq f_1^* + \delta$.

Let us then focus on Case (2). We first notice that the assumed boundedness of $\partial f_1(\mathcal{P}_{X_1}(\mathbf{x}_k))$ leads to

$$f_1(\mathcal{P}_{X_1}(\mathbf{x}_k)) \leq f_1(\mathbf{x}_k) + Md_{X_1}(\mathbf{x}_k).$$

Therefore, $f_1(\mathbf{x}_k) \leq f_1^* + \delta$ implies

$$\mathbf{x}_k \in X_2^{\delta + Md_{X_1}(\mathbf{x}_k)}.$$

Thus, (17) now reads

$$d_{X_2}(\mathbf{x}_{k+1}) \leq \Delta_2(\delta + Md_{X_1}(\mathbf{x}_k)) + \|\mathbf{x}_{k+2/3} - \mathbf{x}_k\|.$$

Then, because we have assumed $d_{X_1}(\mathbf{x}_k) \rightarrow 0$, $\|\mathbf{x}_k - \mathbf{x}_{k+1/3}\| \rightarrow 0$, and $\|\mathbf{x}_{k+1/3} - \mathbf{x}_{k+2/3}\| \rightarrow 0$, we can recall $\lim_{\eta \rightarrow 0} \Delta_2(\delta + \eta) = \Delta_2(\delta)$, so that the argumentation above leads to the conclusion that

$$\limsup_{k \rightarrow \infty} d_{X_2}(\mathbf{x}_k) \leq \Delta_2(\delta).$$

Finally, because $\delta > 0$ was arbitrary and $\lim_{\delta \rightarrow 0} \Delta_2(\delta) = 0$, we have just proven the claimed result. \square

We now present two different algorithms and prove their convergence based on the above general results. Next section contains numerical experimentation regarding some of these methods in four different bilevel models arising in high-resolution micro-tomographic image reconstruction from synchrotron illumination.

2.4 An Algorithm for Lipschitz-Differentiable Primary Objective Functions

In this Subsection we suppose f_0 in the bilevel optimization problem (1) is differentiable with uniformly bounded and Lipschitz continuous gradient. For this kind of problem, we will consider Algorithm 1, which we name Fast Iterative Bilevel Algorithm (FIBA). FIBA first performs a projected gradient descent step, with a stepsize that does not change unless the magnitude of this operation is larger than a control sequence. This computation is then followed by the application of an optimality operator of the kind described by Properties 1 and 2.

Require: $\mathbf{x}_0, \{\lambda_k\}, \{\mu_k\}, \{\zeta_k\}$

1: Initialization: $k \leftarrow 0, t_0 = 1, \mathbf{x}_{-2/3} = \mathbf{x}_0, i_0 = 0$

2: **repeat**

3: $\mathbf{x}_{k+1/3} = \mathcal{P}_{X_0}(\mathbf{x}_k - \lambda_{i_k} \nabla f_0(\mathbf{x}_k))$

4: **if** $\|\mathbf{x}_k - \mathbf{x}_{k+1/3}\| \geq \zeta_k$ **then**

5: $i_{k+1} = i_k + 1$

6: **else**

7: $i_{k+1} = i_k$

8: **end if**

9: $t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \xi_k = \min \left\{ 1, \frac{\mu_k \zeta_k}{\|\mathbf{x}_{k+1/3} - \mathbf{x}_{(k-1)+1/3}\|} \right\}$

10: $\mathbf{y}_{k+1/3} = \mathbf{x}_{k+1/3} + \xi_k \left(\frac{t_k - 1}{t_{k+1}} \right) (\mathbf{x}_{k+1/3} - \mathbf{x}_{(k-1)+1/3})$

11: $\mathbf{x}_{k+2/3} = \mathcal{O}_{f_1}(\mathbf{y}_{k+1/3}, \mu_k)$

12: $\mathbf{x}_{k+1} = \mathcal{P}_{X_0}(\mathbf{x}_{k+2/3})$

13: $k \leftarrow k + 1$

14: **until** convergence is reached

Algorithm 1 Fast Iterative Bilevel Algorithm

Such optimality operator is actually applied to a perturbation of the point obtained by the projected gradient descent, in a fashion similar to the Fast Iterative Soft-Thresholding Algorithm (FISTA) [2], but with the magnitude of the perturbation bounded by $\mu_k \zeta_k$, where $\{\zeta_k\}$ is a positive vanishing sequence.

In order to analyze convergence of Algorithm 1 through our previous results, we first look at the simple projected gradient descent

$$\mathcal{G}_f(\lambda, \mathbf{x}) := \mathcal{P}_{X_0}(\mathbf{x} - \lambda \nabla f(\mathbf{x}))$$

as an instance of the optimality operators considered above. Let L_f denote the Lipschitz constant of ∇f . Then, if $\lambda \leq 1/L_f$, it is possible to show (see, e.g., [2] and references therein) that

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2\lambda} \|\mathbf{y} - \mathbf{x}\|^2. \quad (22)$$

Let now ι_{X_0} be the indicator function:

$$\iota_{X_0}(\mathbf{x}) := \begin{cases} \infty & \text{if } \mathbf{x} \notin X_0 \\ 0 & \text{if } \mathbf{x} \in X_0. \end{cases} \quad (23)$$

Then, inequality (22) leads to

$$f(\mathbf{y}) + \iota_{X_0}(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2\lambda} \|\mathbf{y} - \mathbf{x}\|^2 + \iota_{X_0}(\mathbf{y}).$$

Therefore, [2, Lemma 2.3] can be used with $L = 1/\lambda$, $g = \iota_{X_0}$, $\mathbf{y} = \mathbf{x}$ and $\mathbf{x} = \mathbf{y}$ in order to get, for $\mathbf{y} \in X_0$:

$$\begin{aligned} 2\lambda \left(f(\mathbf{y}) - f(\mathcal{G}_f(\lambda, \mathbf{x})) \right) &\geq \|\mathbf{x} - \mathcal{G}_f(\lambda, \mathbf{x})\|^2 + 2(\mathbf{x} - \mathbf{y})^T (\mathcal{G}_f(\lambda, \mathbf{x}) - \mathbf{x}) \\ &= \|\mathbf{y} - \mathcal{G}_f(\lambda, \mathbf{x})\|^2 - \|\mathbf{y} - \mathbf{x}\|^2. \end{aligned} \quad (24)$$

Thus, for $\lambda \leq 1/L_f$, we can see that $\mathcal{G}_f(\lambda, \mathbf{x})$ satisfies Property 1 with $\beta = 2$ and $\rho(\lambda, k) \equiv 0$. Notice that the operation described at line 3 of Algorithm 1 is actually:

$$\mathbf{x}_{k+1/3} = \mathcal{G}_{f_0}(\lambda_{i_k}, \mathbf{x}_k).$$

Consequently, according to (24), Algorithm 1 is an instance of (9) with an optimality operator \mathcal{O}_{f_0} which satisfies Property 1 with $\rho_0(\lambda, k) \equiv 0$, whenever $\lambda \leq 1/L_f$ or if, which is weaker, (22) holds with $\mathbf{x} = \mathbf{x}_k$ and $\mathbf{y} = \mathbf{x}_{k+1/3}$. This characteristic of the error term implies that it is possible to have a precise enough operator without requiring $\lambda_{i_k} \rightarrow 0$. However, $\|\mathbf{x}_k - \mathbf{x}_{k+1/3}\| \rightarrow 0$ would still require $\lambda_{i_k} \rightarrow 0$ if the projection on line 3 of Algorithm 1 were not performed and this is the reason why there are two projections in this method.

Now, let us recall that $\mathbf{x}_k \in X_0$, so that by the definition of \mathcal{G}_f and (12)

$$\|\mathbf{x}_k - \mathcal{G}_f(\lambda_{i_k}, \mathbf{x}_k)\| \leq \lambda_{i_k} \|\nabla f(\mathbf{x}_k)\|. \quad (25)$$

Therefore, if the sequence $\{\nabla f_0(\mathbf{x}_k)\}$ is bounded, and if $\lambda_k \rightarrow 0$ and $\zeta_k \rightarrow 0$, then the procedure in lines 4-8 of Algorithm 1 implies that $\|\mathbf{x}_{k+1/3} - \mathbf{x}_k\| \rightarrow 0$. Observe also that if $\sum_k \lambda_k = \infty$, then $\sum_k \lambda_{i_k} = \infty$ too.

Now, we consider the fact that the optimization operator for the secondary function f_1 is used in a perturbed point $\mathbf{y}_{k+1/3}$, instead of at $\mathbf{x}_{k+1/3}$. Our goal is to verify that the relevant properties of \mathcal{O}_{f_1} are maintained. The first observation is that, given the way that $\mathbf{y}_{k+1/3}$ is defined, we have

$$\|\mathbf{x}_{k+1/3} - \mathbf{y}_{k+1/3}\| \leq \mu_k \zeta_k. \quad (26)$$

We then define a new operator $\tilde{\mathcal{O}}_{f_1}$, based on \mathcal{O}_{f_1} , as follows:

$$\tilde{\mathcal{O}}_{f_1}(\mu, \mathbf{x}_{k+1/3}) := \mathcal{O}_{f_1}(\mu, \mathbf{y}_{k+1/3}).$$

Notice that $\tilde{\mathcal{O}}_{f_1}$ role is to hide the perturbation from the analysis. Also, this kind of operator is the reason why we use an iteration-dependent error term in Property 1, as we will see just below. Let us then assume that Property 1 holds for \mathcal{O}_{f_1} , therefore:

$$\begin{aligned} \|\mathbf{y} - \tilde{\mathcal{O}}_{f_1}(\mu_k, \mathbf{x}_{k+1/3})\|^2 &= \|\mathbf{y} - \mathcal{O}_{f_1}(\mu_k, \mathbf{y}_{k+1/3})\|^2 \\ &\leq \|\mathbf{y} - \mathbf{y}_{k+1/3}\|^2 - 2\mu_k (f(\mathcal{O}_{f_1}(\mu_k, \mathbf{y}_{k+1/3})) - f(\mathbf{y})) \\ &\quad + \mu_k \rho_1(\mu_k, k) \\ &= \|\mathbf{y} - \mathbf{y}_{k+1/3}\|^2 - 2\mu_k (f(\tilde{\mathcal{O}}_{f_1}(\mu_k, \mathbf{x}_{k+1/3})) - f(\mathbf{y})) \\ &\quad + \mu_k \rho_1(\mu_k, k). \end{aligned} \quad (27)$$

Now, by taking (26) into consideration, a straightforward computation leads to

$$\|\mathbf{y} - \mathbf{y}_{k+1/3}\|^2 \leq \|\mathbf{y} - \mathbf{x}_{k+1/3}\|^2 + \mu_k \zeta_k (2\|\mathbf{y} - \mathbf{x}_{k+1/3}\| + \mu_k \zeta_k).$$

Then, using the above bound in (27) we have:

$$\begin{aligned} \|\mathbf{y} - \tilde{\mathcal{O}}_{f_1}(\mu_k, \mathbf{x}_{k+1/3})\|^2 &\leq \|\mathbf{y} - \mathbf{x}_{k+1/3}\|^2 - 2\mu_k (f(\tilde{\mathcal{O}}_{f_1}(\mu_k, \mathbf{x}_{k+1/3})) - f(\mathbf{y})) \\ &\quad + \mu_k (\rho_1(\mu_k, k) + \zeta_k (2\|\mathbf{y} - \mathbf{x}_{k+1/3}\| + \mu_k \zeta_k)). \end{aligned} \quad (28)$$

That is, $\tilde{\mathcal{O}}_{f_1}$ satisfies Property 1 with ρ_1 replaced by

$$\tilde{\rho}_1(\mu_k, k) := \rho_1(\mu_k, k) + \zeta_k (2\|\mathbf{y} - \mathbf{x}_{k+1/3}\| + \mu_k \zeta_k),$$

where we notice that the set of points \mathbf{y} where Property 1 is applied in the convergence proofs is bounded if $\{\mathbf{x}_k\}$ is bounded.

Given the above considerations, we are ready to provide the convergence results for Algorithm 1.

Theorem 1 *Assume f_0 is differentiable with Lipschitz-continuous gradient and has Lipschitz constant L_0 . Suppose too that f_0 has a bounded gradient and that f_1 has a bounded subgradient and $\{f_1(\mathbf{x}_{k+1/3})\}$ bounded from below. Assume $\{\lambda_k\}$, $\{\mu_k\}$ and $\{\zeta_k\}$ are non-negative vanishing scalar sequences such that $\sum_{k=0}^{\infty} \lambda_k = \infty$, $\lambda_k \leq 1/L_0$ (or each λ_{i_k} satisfies (22)), $\sum_{k=0}^{\infty} \mu_k = \infty$, and $\mu_k/\lambda_k \rightarrow 0$. Then, if $X_2 \neq \emptyset$, $\{\mathbf{x}_k\}$ is bounded, and \mathcal{O}_{f_1} satisfies Properties 1 and 2 with $\rho_1(\mu_k, k) \rightarrow 0$, we have*

$$\lim_{k \rightarrow \infty} d_{X_2}(\mathbf{x}_k) = 0.$$

Proof First let us notice that Algorithm 1 can be written as

$$\begin{aligned} \mathbf{x}_{k+1/3} &:= \mathcal{G}_{f_0}(\tilde{\lambda}_k, \mathbf{x}_k); \\ \mathbf{x}_{k+2/3} &:= \tilde{\mathcal{O}}_{f_1}(\mu_k, \mathbf{x}_{k+1/3}); \\ \mathbf{x}_{k+1} &:= \mathcal{P}_{X_0}(\mathbf{x}_{k+2/3}), \end{aligned}$$

where $\tilde{\lambda}_k := \lambda_{i_k}$. Since the construction of the algorithm guarantees that $i_k \leq k$, we have $\sum_{k=0}^{\infty} \tilde{\lambda}_k = \infty$ and $\mu_k/\tilde{\lambda}_k \rightarrow 0$. Because $\tilde{\lambda}_k \leq 1/L_0$, (24) holds and therefore, \mathcal{G}_{f_0} satisfies Property 1 with $\rho_0(\tilde{\lambda}_k, k) \equiv 0$. Furthermore, because (25) and the algorithm definition, as already argued, we have $\|\mathbf{x}_k - \mathbf{x}_{k+1/3}\| \rightarrow 0$. Also, if \mathcal{O}_{f_1} satisfies Property 2 so does $\tilde{\mathcal{O}}_{f_1}$. Furthermore, as shown above, if \mathcal{O}_{f_1} satisfies Property 1 so does $\tilde{\mathcal{O}}_{f_1}$, with the error term given by the factor multiplying μ_k in the second line of (28). Thus, the assumed boundedness of $\{\mathbf{x}_k\}$ and of f_1 ensure that Proposition 2 can be applied so that

$$\lim_{k \rightarrow \infty} d_{X_1}(\mathbf{x}_k) = 0.$$

Now, because $\mathbf{x}_{k+1/3} \in X_0$, we then have $f_0(\mathbf{x}_{k+1/3}) \geq f_0^*$. Furthermore, because of the boundedness assumptions and of $\mu_k \rightarrow 0$, it is possible to see that $\tilde{\rho}_1(\mu_k, k) \rightarrow 0$. Therefore, Proposition 3 can be applied, which leads to the desired conclusion. \square

2.5 Incremental Algorithms for Non-Differentiable Problems

Here we specialize Algorithm (9) to the case where f_0 is the sum of several non-differentiable convex functions:

$$f_0 := \sum_{i=1}^m f_0^i.$$

In this situation we propose the use, for the primary optimization problem, of the incremental subgradient operator, denoted as $\mathcal{I}_f : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$, given by:

$$\begin{aligned} \mathbf{x}^{(1)} &= \mathbf{x} \\ \mathbf{x}^{(i+1)} &= \mathbf{x}^{(i)} - \lambda \tilde{\nabla} f^i(\mathbf{x}^{(i)}) \quad i = 1, 2, \dots, m \\ \mathcal{I}_f(\lambda, \mathbf{x}) &= \mathbf{x}^{(m+1)}. \end{aligned}$$

Incremental operators are well known for its fast initial convergence rate and, accordingly, several variations of it have been thoroughly analyzed in the literature [4–6, 13, 29, 36, 37]. We will use here the result [29, Lemma 2.1]:

Lemma 1 *Assume the subgradients of the convex functions f_0^i are bounded in the following sense:*

$$\forall \mathbf{x} \in \mathbb{R}^n \quad \text{and} \quad \forall \mathbf{v} \in \partial f_0^i(\mathbf{x}), \quad \|\mathbf{v}\| \leq C_i, \quad i \in \{1, 2, \dots, m\}. \quad (29)$$

Then, the incremental subgradient operator satisfies, for every $\lambda \in \mathbb{R}_+$, and $\mathbf{y}, \mathbf{x} \in \mathbb{R}^n$:

$$\|\mathcal{I}_{f_0}(\lambda, \mathbf{x}) - \mathbf{y}\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 - 2\lambda(f_0(\mathbf{x}) - f_0(\mathbf{y})) + \lambda^2 \left(\sum_{i=1}^m C_i \right)^2, \quad (30)$$

where $f_0 := \sum_{i=1}^m f_0^i$.

Now, notice that the boundedness condition on the subdifferentials leads, for every $\mathbf{x} \in \mathbb{R}^n$, to

$$\|\mathcal{I}_{f_0}(\lambda, \mathbf{x}) - \mathbf{x}\| \leq \lambda \sum_{i=1}^m C_i, \quad \text{and} \quad \tilde{\nabla} f_0(\mathbf{x}) \in \partial f_0(\mathbf{x}) \Rightarrow \|\nabla f_0(\mathbf{x})\| \leq \sum_{i=1}^m C_i. \quad (31)$$

Thus, applying Proposition 1 we are lead to the following result:

Corollary 1 *Assume the subgradients of the convex functions f_0^i satisfy (29). Then, the incremental subgradient operator satisfies, for every $\lambda \in \mathbb{R}_+$, and $\mathbf{y}, \mathbf{x} \in \mathbb{R}^n$:*

$$\|\mathcal{I}_{f_0}(\lambda, \mathbf{x}) - \mathbf{y}\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 - 2\lambda(f_0(\mathcal{I}_{f_0}(\lambda, \mathbf{x})) - f_0(\mathbf{y})) + 3\lambda^2 \left(\sum_{i=1}^m C_i \right)^2,$$

where, again, $f_0 := \sum_{i=1}^m f_0^i$.

Require: $\mathbf{x}_0, \{\lambda_k\}, \{\mu_k\}$

1: Initialization: $k \leftarrow 0$

2: **repeat**

3: $\mathbf{x}_{k+1/3} = \mathcal{I}_{f_0}(\mathbf{x}_k)$

4: $\mathbf{x}_{k+2/3} = \mathcal{O}_{f_1}(\mathbf{x}_{k+1/3}, \mu_k)$

5: $\mathbf{x}_{k+1} = \mathcal{P}_{X_0}(\mathbf{x}_{k+2/3})$

6: $k \leftarrow k + 1$

7: **until** convergence is reached

Algorithm 2 Incremental Iterative Bilevel Algorithm

The second algorithm we propose in this work will be called IIBA, from Incremental Iterative Bilevel Algorithm, and is described in Algorithm 2 below. For this algorithm we have the following convergence result.

Theorem 2 Assume that f_0 is of the form $f_0 := \sum_{i=1}^m f_0^i$ and satisfies (29), that f_1 has a bounded subgradient, and that $\{f_1(\mathbf{x}_{k+1/3})\}$ is bounded from below. Assume $\{\lambda_k\}$ and $\{\mu_k\}$ are non-negative vanishing scalar sequences such that $\sum_{k=0}^{\infty} \lambda_k = \infty$, $\sum_{k=0}^{\infty} \mu_k = \infty$, $\mu_k/\lambda_k \rightarrow 0$ and $\lambda_k^2/\mu_k \rightarrow 0$. Then, suppose $X_2 \neq \emptyset$ and X_2 is bounded (or $\{\mathbf{x}_k\}$ is bounded), and \mathcal{O}_{f_1} satisfies Properties 1 and 2 with $\rho_1(\mu_k, k) \rightarrow 0$. Then, the sequence $\{\mathbf{x}_k\}$ generated by Algorithm 2 satisfies

$$\lim_{k \rightarrow \infty} d_{X_2}(\mathbf{x}_k) = 0.$$

Proof Notice that Lemma 1 together with Proposition 1 and the subgradient boundedness assumption imply that \mathcal{I}_{f_0} satisfies the desired Property 1. Also, $\lambda_k \rightarrow 0$ implies, together with the subgradient boundedness assumption, that $\|\mathbf{x}_k - \mathbf{x}_{k+1/3}\| \rightarrow 0$ and that $\rho_0(\lambda_k, k) \rightarrow 0$, the latter because of Lemma 1. Therefore, Proposition 2 implies that

$$d_{X_1}(\mathbf{x}_k) \rightarrow 0.$$

Now, notice the subgradient boundedness assumption and (31) imply that $[f_0(\mathbf{x}_k) - f_0(\mathbf{x}_{k+1/3})]_+ = O(\lambda_k)$, and, therefore, since $\mathbf{x}_k \in X_0$ for $k > 0$, we have $[f_0^* - f(\mathbf{x}_{k+1/3})]_+ = O(\lambda_k)$. Thus, $\lambda_k^2/\mu_k \rightarrow 0$ implies $\lambda_k[f_0^* - f(\mathbf{x}_{k+1/3})]_+/\mu_k \rightarrow 0$. Furthermore, since, by (30), $\rho_0(\lambda_k, k) = O(\lambda_k)$, λ_k^2/μ_k also implies $\lambda_k \rho_0(\lambda_k, k)/\mu_k \rightarrow 0$. So, finally, Proposition 3 can be applied, which proves the result. \square

2.6 Stopping Criterion

Here we devise a stopping criterion for the proposed bilevel methods, based on inequalities (15) and (20) coupled to the following Lemma:

Lemma 2 Let $\{a_k\}$ and $\{b_k\}$ be non-negative sequences such that $a_k/b_k \rightarrow 0$ and $\sum_{k=0}^{\infty} b_k = \infty$. Then

$$\lim_{n \rightarrow \infty} \frac{\sum_{k=0}^n a_k}{\sum_{k=0}^n b_k} = 0.$$

Proof Choose any $\alpha > 0$ and let k_0 be such that $a_k/b_k \leq \alpha$ for every $k \geq k_0$. Then

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\sum_{k=0}^n a_k}{\sum_{k=0}^n b_k} &= \lim_{n \rightarrow \infty} \frac{\sum_{k=0}^{k_0-1} a_k + \sum_{k=k_0}^n a_k}{\sum_{k=0}^n b_k} \\ &= \lim_{n \rightarrow \infty} \frac{\sum_{k=k_0}^n a_k}{\sum_{k=0}^n b_k}. \end{aligned}$$

But

$$\frac{\sum_{k=k_0}^n a_k}{\sum_{k=0}^n b_k} \leq \frac{\sum_{k=k_0}^n a_k}{\sum_{k=k_0}^n b_k} = \frac{\sum_{k=k_0}^n (a_k/b_k) b_k}{\sum_{k=k_0}^n b_k} \leq \alpha.$$

Therefore, $\lim_{n \rightarrow \infty} |\sum_{k=0}^n a_k / \sum_{k=0}^n b_k| \leq \alpha$ for any $\alpha > 0$. \square

In order to explain the stopping criterion, we resort to the concept of best-so-far iteration. Let us denote as $\phi_i^{k_0, k}$, for $i \in \{0, 1\}$ and $k_0 \leq k$ integers in $\{0, 1, \dots\}$, the smallest value in the set $\{f_i(\mathbf{x}_{k_0}), f_i(\mathbf{x}_1), \dots, f_i(\mathbf{x}_k)\}$. For the special case $k_0 = 0$, we simplify the notation by $\phi_i^k := \phi_i^{0, k}$. Then, successive application of (15) together with $\phi_i^k \leq f_i(\mathbf{x}_j)$ for all $j \leq k$ leads to

$$\begin{aligned} \phi_0^k - f_0^* &\leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{\beta \sum_{i=0}^k \lambda_i} + \frac{\sum_{i=0}^k \lambda_i (\rho_0(\lambda_i, i) + \beta M \|\mathbf{x}_{i+1/3} - \mathbf{x}_i\|)}{\beta \sum_{i=0}^k \lambda_i} \\ &\quad + \frac{\sum_{i=0}^k \mu_i (\beta N + \rho_1(\mu_i, i))}{\beta \sum_{i=0}^k \lambda_i} =: \sigma_0^k. \end{aligned}$$

Therefore, under the hypothesis of Proposition 2, Lemma 2 ensures that the right-hand side of the above inequality vanishes as the iterations proceed. Thus, the quantity σ_0^k can be used as a measure of convergence to X_1 bounding the difference between the best f_0 function value to the optimal f_0^* , as long as it is possible to estimate the distance $\|\mathbf{x}_0 - \mathbf{x}^*\|$, the behavior of the error terms and the subgradient bounding constants. Notice that, in principle, the constant N require knowledge of the optimal value in this case, but it can be replaced by an upper bound for it, which should not be difficult to obtain in many cases.

We can use a very similar reasoning in order to estimate optimality of the secondary objective function too. Applying (20) repeatedly and recalling that $\phi_1^{k_0, k} \leq f_1(\mathbf{x}_i)$ for every $i \in \{k_0, k_0 + 1, \dots, k\}$ we have

$$\begin{aligned} \phi_1^{k_0, k} - f_1^* &\leq \frac{\|\mathbf{x}_{k_0} - \mathbf{x}^*\|^2}{\beta \sum_{i=k_0}^k \mu_i} + \frac{\sum_{i=k_0}^k \lambda_i (\rho_0(\lambda_i, i) + \beta [f_0^* - f_0(\mathbf{x}_{k+1/3})]_+)}{\beta \sum_{i=k_0}^k \mu_i} \\ &\quad + \frac{\sum_{i=k_0}^k \mu_i (\rho_1(\mu_i, i) + \beta M \|\mathbf{x}_{i+2/3} - \mathbf{x}_i\|)}{\beta \sum_{i=k_0}^k \mu_i} =: \sigma_1^{k_0, k}. \end{aligned}$$

Now, under the hypothesis of Proposition 3, Lemma 2 ensures that for any k_0 , we have $\lim_{k \rightarrow \infty} \sigma_0^{k_0, k} = 0$.

We can now describe how to stop the algorithm at a non-negative integer $\kappa_{\epsilon_0, \epsilon_1}$ such that we have

$$f_0(\mathbf{x}_{\kappa_{\epsilon_0, \epsilon_1}}) - f_0^* \leq \epsilon_0 \quad \text{and} \quad f_1(\mathbf{x}_{\kappa_{\epsilon_0, \epsilon_1}}) - f_1^* \leq \epsilon_1,$$

for any pair of positive numbers ϵ_0 and ϵ_1 . Let us consider the following procedure:

1. Iterate the algorithm until $\sigma_0^k \leq \epsilon_0$;
2. $k_0 \leftarrow k$, $\kappa \leftarrow k$;
3. Iterate the algorithm until $\sigma_1^{\kappa, k} \leq \epsilon_1$;
4. Let $k_1 \geq \kappa$ be such that $f_1(\mathbf{x}_{k_1}) = \phi_1^{\kappa, k}$;
5. If $f_0(\mathbf{x}_{k_1}) \leq \phi_0^{k_0}$: STOP;
6. $\kappa \leftarrow k$; Go to step 3.

Notice that once the procedure has stopped, then

$$f_0(\mathbf{x}_{k_1}) - f_0^* \leq \phi_0^{k_0} - f_0^* \leq \sigma_0^k \leq \epsilon_0,$$

and

$$f_1(\mathbf{x}_{k_1}) - f_1^* = \phi_1^{\kappa, k} - f_1^* \leq \sigma_1^{\kappa, k} \leq \epsilon_1.$$

The procedure indeed stops because since Proposition 2 ensures convergence in norm to X_1 , mild subgradient boundedness assumptions will guarantee also that $f_0(\mathbf{x}_k) \rightarrow f_0^*$. Therefore we know that for large enough k , there holds $f_0(\mathbf{x}_k) \leq \phi_0^{k_0}$. That is, the condition in step 5 will eventually be satisfied as the iterations proceed.

3 Application Problem Presentation

We have performed experiments involving tomographic image reconstruction. In tomography, the idealized problem is to reconstruct a function $\mu : \mathbb{R}^2 \rightarrow \mathbb{R}$ given the values of its integrals along straight lines, that is, given its *Radon transform* denoted as $\mathcal{R}[\mu]$ and defined by the following equality:

$$\mathcal{R}[\mu](\theta, t) := \int_{\mathbb{R}} \mu \left(t \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix} + s \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix} \right) ds.$$

Figure 1 (adapted from [21]) brings a graphical representation of this definition.

Data was obtained by transmitting synchrotron radiation through samples of eggs taken from a fish of the species *Prochilodus lineatus* collected at the Madeira River's bed, immersed in distilled water inside a capillary test tube. Data acquisition was performed at the Brazilian National Synchrotron Light Source (LNLS)¹.

¹ <http://lnls.cnpem.br/>

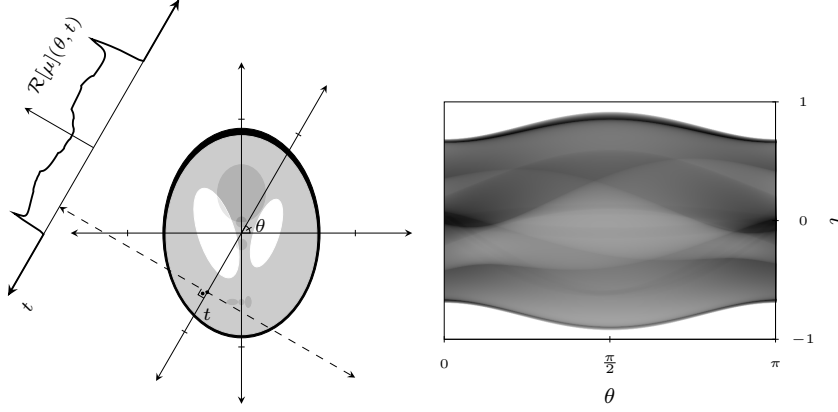


Fig. 1 Left: schematic representation of the Radon transform. In the definition, θ is the angle between the normal to the integration path and the horizontal axis, while t is the line of integration's displacement from the origin. Right: Image of the Radon transform of the image shown on the left in the $\theta \times t$ coordinate system.

In a transmission tomography setup [23, 25, 28] like the one we have used, the value of the line integral is estimated through the emitted to detected intensity ratio according to Beer–Lambert law:

$$\frac{I_e}{I_d} = e^{\int_L \mu(\mathbf{x}) ds},$$

where I_e is the emitted intensity, I_d is the detected intensity, μ gives the linear attenuation factor of the imaged object at each point in space, and L is the straight line connecting detector to emitter. While in this case the reconstruction problem is essentially bi-dimensional, a simultaneous acquisition of a radiography of 2048 parallel slices of the object to be imaged is made at each angle, which enables volumetric reconstruction, if desired, by the stacking of several bi-dimensional reconstructions. After each plain x-ray imaging, the sample is rotated and new samples of the Radon transform are estimated in the same manner at a new angle. Figure 2 depicts the process of assembling the 2048×200 data array, which will be used for a slice reconstruction, from the 200 images of size 2048×2048 .

In our application the imaged subject is sensitive in a way such that a overly long exposure time under a low energy x-ray beam may overheat or otherwise physically damage the sample. Therefore, because the exposure time for good radiographies under a monochromatic beam at LNLS' facilities was experimentally found to be at least 20 seconds, the Radon Transform was sampled at only 200 evenly spaced angles covering the interval $[-\pi, 0]$, a relatively small number if we are willing to reconstruct full resolution 2048×2048 images from this data. In this case, it is likely that problem (34) will have many solutions and we need to select one of these, therefore the need of a bilevel model arises.

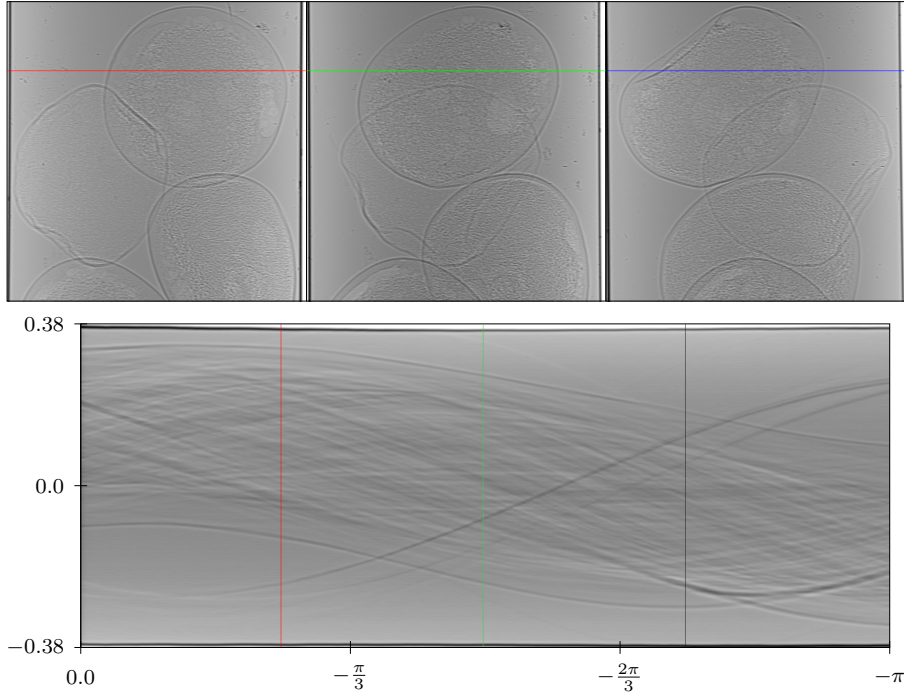


Fig. 2 Assembly of the Radon Transform. Top: three of the 200 radiographic images used, each of which has 2048×2048 pixels and depicts a square region of area $0.76 \times 0.76 \text{mm}^2$. Bottom: the n^{th} row of the i^{th} image has samples of $\mathcal{R}[g_n](\theta_i, \cdot)$, that is, a column in the representation of $\mathcal{R}[g_n]$ in the $\theta \times t$ plane, where $g_n : \mathbb{R}^2 \rightarrow \mathbb{R}$ gives the linear attenuation factor at each point in the n^{th} slice to be reconstructed. The colored solid lines depict the position of the radiographies' rows in the resulting sinogram.

3.1 Primary Objective Functions

Assuming the original image $g : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ lies in a finite dimensional vector space generated by some basis $\{g^1, g^2, \dots, g^n\}$ and recognizing that the number of measurements is always finite in practice, one can reduce the problem of tomographic reconstruction to a linear system of equations:

$$R\mathbf{x} = \mathbf{b}, \quad (32)$$

where the elements r_{ij} of the matrix R are given by

$$\mathcal{R}[g^j](\theta_i, t_i) \quad (33)$$

and the elements b_i of the vector \mathbf{b} are the corresponding experimental data, that is, b_i is an approximate sample of $\mathcal{R}[g](\theta_i, t_i)$, where $g = \sum_{i=1}^n x_i g^i$ is the desired image. Because actual microtomographic data from synchrotron illumination will always contain errors (either from micrometric misalignment of the experimental setup, small dust particles in the emitter-detector path, or from statistical fluctuation of photon emission and attenuation), the above linear system of equations may not have a solution.

3.2 Differentiable Primary Objective Function

System (33) can be replaced by a constrained least squares problem, in order to alleviate the likely lack of consistency:

$$X_1 = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}_+^n} q(\mathbf{x}) := \frac{1}{2} \|R\mathbf{x} - \mathbf{b}\|^2. \quad (34)$$

Therefore, in this case we will be lead to a bilevel problem of the form (1) with $X_0 = \mathbb{R}_+^n$.

Another option for a feasibility set would be the bounded box $\{\mathbf{x} \in \mathbb{R}^n : 0 \leq x_i \leq u\}$ for some $u > 0$. This may be a sensible idea if the maximum attenuation factor is known beforehand and we could thus include this information into the model. Because we do not make this kind of boundedness imposition, by using the least squares function we are at the risk of violating certain (sub)gradient boundedness assumption. We therefore use another continuously differentiable convex function f_0 , which has an uniformly bounded gradient and is such that

$$\operatorname{argmin}_{\mathbf{x} \in X_0} f_0(\mathbf{x}) = \operatorname{argmin}_{\mathbf{x} \in X_0} q(\mathbf{x}). \quad (35)$$

In order to build such function, take any $\tilde{\mathbf{x}} \in X_0$ (such as, in our case, $\tilde{\mathbf{x}} = \mathbf{0}$) and set

$$\Delta := \|R\tilde{\mathbf{x}} - \mathbf{b}\| > 0,$$

then define

$$f_0(\mathbf{x}) := \frac{1}{2} \sum_{i=1}^m h(\langle r_i, \mathbf{x} \rangle - b_i),$$

where r_i^T is the i -th row of R and the function $h : \mathbb{R} \rightarrow \mathbb{R}_+$ is given as

$$h(x) := \begin{cases} x^2 & \text{if } |x| < \Delta \\ 2\Delta|x| - \Delta^2 & \text{otherwise.} \end{cases}$$

Notice that if $q(\mathbf{x}) \leq q(\tilde{\mathbf{x}}) = 1/2\Delta^2$, then $f_0(\mathbf{x}) = q(\mathbf{x})$ and, furthermore, if $f_0(\mathbf{x}) \neq q(\mathbf{x})$, then $f_0(\mathbf{x}) \geq 1/2\Delta^2 = q(\tilde{\mathbf{x}})$. Therefore, (35) holds.

We need to show that the uniformly bounded continuous differentiability and convexity claims hold, but these facts follow from convexity and continuous differentiability of h . Its derivative can be computed to be

$$h'(x) := \begin{cases} 2x & \text{if } |x| < \Delta \\ 2\Delta \operatorname{sign}(x) & \text{otherwise,} \end{cases}$$

with $\operatorname{sign} : \mathbb{R} \rightarrow 2^{\mathbb{R}}$ defined as

$$\operatorname{sign}(x) := \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0. \end{cases}$$

Thus, the derivative of h is uniformly bounded by $|h'(x)| \leq 2\Delta$, is continuous as long as we use $\Delta > 0$, and, since $h'(x)$ is nondecreasing, h is convex. Notice that h is the well known Huber function [24].

In order to simplify the notation, we introduce the function $\mathbf{h}' : \mathbb{R}^m \rightarrow \mathbb{R}^m$ given as:

$$\mathbf{h}'(\mathbf{x}) := \begin{pmatrix} h'(x_1) \\ h'(x_2) \\ \vdots \\ h'(x_m) \end{pmatrix}.$$

Under this notation, a straightforward calculation leads to

$$\nabla f_0(\mathbf{x}) = \frac{1}{2} R^T \mathbf{h}'(R\mathbf{x} - \mathbf{b}).$$

The computationally expensive parts of this expression are products of the form $R\mathbf{x}$ and $R^T \mathbf{b}$. In our case, because of image resolution and dataset size, as explained in the previous Subsection, matrix R has dimensions of over $4 \cdot 10^5$ lines by $4 \cdot 10^6$ columns.

In order to obtain a computationally viable algorithm, these matrix-vector products must be amenable to fast computation, and, in fact, there is a very effective approach to it based in a relation between the one-dimensional Fourier Transform of the data and one “slice” of the two-dimensional Fourier Transform of the original image. Recall that we denote our desired, unknown, image as $g : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ and define the projection of f at an angle t as

$$p_\theta(t) := \mathcal{R}[g](t).$$

Now, using the hat notation for the Fourier transform of an integrable function $f : \mathbb{R}^n \rightarrow \mathbb{C}$, as defined as follows, where $j := \sqrt{-1}$:

$$\hat{f}(\boldsymbol{\omega}) := \mathcal{F}[f](\boldsymbol{\omega}) := \int_{\mathbb{R}^n} f(\mathbf{x}) e^{-j2\pi \langle \boldsymbol{\omega}, \mathbf{x} \rangle} d\mathbf{x},$$

there holds the so called Fourier-slice theorem [25, 28]:

$$\hat{p}_\theta(\omega) = \hat{g}\left(\omega \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}\right).$$

This result implies that the action of R can be evaluated first in the Fourier space, an operation which can be computed efficiently by Non-uniform Fast Fourier Transforms (NFFT) [18] and later translated back to the original feature space by regular Fast Fourier Transforms (FFT). We have used the `pynfft` binding for the `nfft3` library [26] in order to compute samples of \hat{p}_θ following (3.2), and then used regular inverse FFT routines from `numpy` for the final computation from the Fourier back to the feature space. Fast evaluation of the transpose operation is immediately available from the same set of libraries using the FFT from `numpy` and the transpose of the NFFT available from `nfft3`.

3.3 Non-Differentiable Primary Objective Function

Another option to circumvent the non-consistency of (33) under non-negativity constraints is to use the least 1-norm approach, which is useful because synchrotron illuminated tomographic data contains mostly very mild noise, but also has sparsely distributed highly perturbed points for example caused by small dust particles and other detector failures such as those caused by defective scintillator crystals and other mechanical, thermal or electronic causes [27].

Therefore, our second option for primary optimization problem was given by

$$X_1 = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}_+^n} \ell(\mathbf{x}) := \|\mathbf{R}\mathbf{x} - \mathbf{b}\|_1. \quad (36)$$

Where, now, the primary objective function naturally has an everywhere bounded subdifferential and, as is well known, the $\|\cdot\|_1$ is more forgiving to sparse noise than is the euclidean norm [11, 14, 32, 38].

3.4 Secondary Objective Functions

For the purpose of selecting one among the solutions of (34) or (36), we will consider two particular cases of the bilevel program (1).

3.4.1 Haar 1-Norm

Here the secondary objective function f_1 is given by

$$f_{\text{Haar}}(\mathbf{x}) := \|\mathbf{H}\mathbf{x}\|_1,$$

with $\mathbf{H} \in \mathbb{R}^{n \times n}$ orthonormal, that is, satisfies $\mathbf{H}^T \mathbf{H} = \mathbf{I}$, where \mathbf{I} is the identity matrix and superscript T indicates transposition. Transformation \mathbf{H} is usually a sparsefying transform, in our case the Haar transform, and one is looking for a sparse (in some cases the sparsest [9, 10, 15–17]), in the \mathbf{H} -transformed space, optimizer of f_0 over X_0 .

Notice that for this particular function we have

$$\partial f_{\text{Haar}}(\mathbf{x}) = \mathbf{H}^T \mathbf{sign}(\mathbf{H}\mathbf{x}), \quad (37)$$

where the set-valued function $\mathbf{sign} : \mathbb{R}^n \rightarrow 2^{\mathbb{R}^n}$ is given by

$$\mathbf{sign}(\mathbf{v}) := \{\mathbf{x} \in \mathbb{R}^n : x_i \in \text{sign}(v_i)\}.$$

Let us now consider the soft-thresholding operator, given componentwise as

$$(\text{ST}_\mu(\mathbf{x}))_i := \begin{cases} x_i + \mu & \text{if } x_i < -\mu \\ 0 & \text{if } x_i \in [-\mu, \mu] \\ x_i - \mu & \text{if } x_i > \mu. \end{cases}$$

We then define an intermediary optimality operator as

$$\mathcal{N}_{f_{\text{Haar}}}(\mu, \mathbf{x}) := H^T \text{ST}_\mu(H\mathbf{x}). \quad (38)$$

A straightforward computation convinces us that

$$\mathcal{N}_{f_{\text{Haar}}}(\mu, \mathbf{x}) = \mathbf{x} - \mu \tilde{\nabla} f_1(\bar{\mathbf{x}}), \quad (39)$$

where

$$(H\bar{\mathbf{x}})_i = \begin{cases} (H\mathbf{x})_i - \text{sign}((H\mathbf{x})_i)\mu & \text{if } |(H\mathbf{x})_i| > \mu \\ 0 & \text{if } |(H\mathbf{x})_i| \leq \mu \end{cases}.$$

This means that $\bar{\mathbf{x}} = \mathcal{N}_{f_{\text{Haar}}}(\mu, \mathbf{x})$, and we have denoted $\tilde{\nabla} f_1(\bar{\mathbf{x}}) \in \partial f_{\text{Haar}}(\bar{\mathbf{x}})$. Now we proceed:

$$\begin{aligned} \|\mathcal{N}_{f_{\text{Haar}}}(\mu, \mathbf{x}) - \mathbf{y}\|^2 &= \|\mathbf{x} - \mu \tilde{\nabla} f_{\text{Haar}}(\bar{\mathbf{x}}) - \mathbf{y}\|^2 \\ &= \|\mathbf{x} - \mathbf{y}\|^2 - 2\mu \langle \tilde{\nabla} f_{\text{Haar}}(\bar{\mathbf{x}}), \mathbf{x} - \mathbf{y} \rangle + \|\mu \tilde{\nabla} f_{\text{Haar}}(\bar{\mathbf{x}})\|^2 \\ &= \|\mathbf{x} - \mathbf{y}\|^2 - 2\mu \langle \tilde{\nabla} f_{\text{Haar}}(\bar{\mathbf{x}}), \bar{\mathbf{x}} - \mathbf{y} \rangle \\ &\quad - 2\mu \langle \tilde{\nabla} f_{\text{Haar}}(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle + \|\mu \tilde{\nabla} f_{\text{Haar}}(\bar{\mathbf{x}})\|^2 \end{aligned} \quad (40)$$

and we then can observe from (37) that ∂f_1 is bounded in a way such that $\|\tilde{\nabla} f_1(\mathbf{x})\| \leq \sqrt{n}$ for every \mathbf{x} , and from (40) that $\|\mathbf{x} - \bar{\mathbf{x}}\| \leq \mu\sqrt{n}$ for every \mathbf{x} . Using this facts followed by the subgradient inequality we have

$$\begin{aligned} \|\mathcal{N}_{f_{\text{Haar}}}(\mu, \mathbf{x}) - \mathbf{y}\|^2 &\leq \|\mathbf{x} - \mathbf{y}\|^2 - 2\mu \langle \tilde{\nabla} f_{\text{Haar}}(\bar{\mathbf{x}}), \bar{\mathbf{x}} - \mathbf{y} \rangle + 3\mu^2 n \\ &\leq \|\mathbf{x} - \mathbf{y}\|^2 - 2\mu (f_{\text{Haar}}(\bar{\mathbf{x}}) - f_1(\mathbf{y})) + 3\mu^2 n. \end{aligned} \quad (41)$$

From where we see that Property 1 is satisfied with $\rho_1(\mu, k) = 3\mu n$. Furthermore, since for every \mathbf{x} we have $\|\tilde{\nabla} f_{\text{Haar}}(\mathbf{x})\| \leq \sqrt{n}$, formulation (39) of \mathcal{O}_{f_1} implies that this operator satisfies Property 2 with $\gamma = \sqrt{n}$.

It is worth remarking that the soft-thresholding is an example of a proximal operator,

$$\mathcal{N}_{f_{\text{Haar}}}(\mu, \mathbf{x}) = \underset{\mathbf{y} \in \mathbb{R}^n}{\text{argmin}} \left\{ \mu \|H\mathbf{y}\|_1 + \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2 \right\}.$$

In fact, it is possible to carry out an argumentation similar to the one that lead to inequality (41) above for more general proximal operators, because they too can be interpreted as examples of implicit subgradient algorithms or, alternatively, as appropriately controlled ϵ -subgradient methods [12].

3.4.2 Total Variation

Another function we have used to select among the primary problem optimizers is the Total Variation [31]:

$$f_{\text{TV}}(\mathbf{x}) := \sum_{i=1}^{\sqrt{n}} \sum_{j=1}^{\sqrt{n}} \sqrt{(x_{i,j} - x_{i-1,j})^2 + (x_{i,j} - x_{i,j-1})^2},$$

where we have assumed that n is a perfect square, which is true for our reconstructed images, the lexicographical notation:

$$x_{i,j} := x_{i+(j-1)\sqrt{n}}, \quad (i,j) \in \{1, 2, \dots, \sqrt{n}\}^2$$

and also the following boundary conditions:

$$x_{0,j} = x_{\sqrt{n},j}, \quad x_{i,0} = x_{i,\sqrt{n}}, \quad (i,j) \in \{1, 2, \dots, \sqrt{n}\}^2.$$

The f_{TV} subdifferential is bounded and, therefore, application of a subgradient step would satisfy the required operator properties. However given the low computational time of the f_{TV} subgradient step, it is reasonable to apply it several times instead of a single pass. In this case we can see that the resulting operation would still satisfy the required properties, as follows. Let us denote by $\hat{\mathcal{O}}_f^J$ the following operator:

$$\begin{aligned} \mathbf{x}^{(0)} &:= \mathbf{x}; \\ \mathbf{x}^{(i)} &:= \tilde{\mathcal{O}}_f(\lambda/i, \mathbf{x}^{(i-1)}), \quad i \in \{1, 2, \dots, J\}; \\ \hat{\mathcal{O}}_f^J(\lambda, \mathbf{x}) &:= \mathbf{x}^{(J)}. \end{aligned} \tag{42}$$

That is, the J -fold repetition of $\tilde{\mathcal{O}}_f$, with diminishing stepsizes. For this kind of operator there holds the following:

Lemma 3 *Suppose that f has a bounded subgradient and that $\tilde{\mathcal{O}}_f$ satisfies Properties 1 and 2. Then $\hat{\mathcal{O}}_f^J$ also satisfies Properties 1 and 2.*

Proof Repeated use of Property 2 and the triangle inequality leads to

$$\|\hat{\mathcal{O}}_f^J(\lambda, \mathbf{x}) - \mathbf{x}\| \leq \lambda\gamma \sum_{i=1}^J \frac{1}{i}.$$

More generally, telescoping from j to l we have

$$\|\mathbf{x}^{(j)} - \mathbf{x}^{(l)}\| \leq \lambda\gamma \sum_{i=j+1}^l \frac{1}{i}. \tag{43}$$

Now, repeated use of Property 1 gives

$$\begin{aligned}\|\hat{\mathcal{O}}_f^J(\lambda, \mathbf{x}) - \mathbf{y}\|^2 &\leq \|\mathbf{x} - \mathbf{y}\|^2 - 2\lambda \sum_{i=1}^J \frac{1}{i} (f(\mathbf{x}^{(i)}) - f(\mathbf{y})) + \lambda \sum_{i=1}^J \frac{1}{i} \rho(\lambda/i, k) \\ &= \|\mathbf{x} - \mathbf{y}\|^2 - 2\lambda \sum_{i=1}^J \frac{1}{i} (f(\mathbf{x}^{(J)}) - f(\mathbf{y})) + \lambda \sum_{i=1}^J \frac{1}{i} \rho(\lambda/i, k) \\ &\quad + 2\lambda \sum_{i=1}^J \frac{1}{i} (f(\mathbf{x}^{(J)}) - f(\mathbf{x}^{(i)})).\end{aligned}$$

Then, by subgradient boundedness and from (43), we have:

$$\begin{aligned}\|\hat{\mathcal{O}}_f^J(\lambda, \mathbf{x}) - \mathbf{y}\|^2 &\leq \|\mathbf{x} - \mathbf{y}\|^2 - 2\lambda \sum_{i=1}^J \frac{1}{i} (f(\mathbf{x}^{(J)}) - f(\mathbf{y})) + \lambda \sum_{i=1}^J \frac{1}{i} \rho(\lambda/i, k) \\ &\quad + 2\lambda M \sum_{i=1}^J \frac{1}{i} \|\mathbf{x}^{(J)} - \mathbf{x}^{(i)}\| \\ &\leq \|\mathbf{x} - \mathbf{y}\|^2 - 2\lambda \sum_{i=1}^J \frac{1}{i} (f(\mathbf{x}^{(J)}) - f(\mathbf{y})) + \lambda \sum_{i=1}^J \frac{1}{i} \rho(\lambda/i, k) \\ &\quad + 2\lambda^2 \gamma M \sum_{i=1}^J \frac{1}{i} \sum_{j=i+1}^J \frac{1}{j}.\end{aligned}$$

Therefore, $\hat{\mathcal{O}}_f^J$ also satisfies Property 2, with error term given by

$$\sum_{i=1}^J \frac{1}{i} \rho(\lambda/i, k) + 2\lambda \gamma M \sum_{i=1}^J \frac{1}{i} \sum_{j=i+1}^J \frac{1}{j}. \square$$

Notice that if the original error term $\rho(\lambda, k)$ for $\tilde{\mathcal{O}}_f$ is $O(\lambda)$, then so is the one of the iterated operator $\hat{\mathcal{O}}_f^J$. This remark will be useful when considering convergence of the incremental algorithm to be presented in Subsection 2.5.

4 Numerical Experimentation

4.1 Smooth Primary Objective Function

In the present Subsection, we report the results obtained using Algorithm 1 in order to approximately solve the optimization problem

$$\begin{aligned}\min \quad & f_1(\mathbf{x}) \\ \text{s. t.} \quad & \mathbf{x} \in \underset{\mathbf{y} \in \mathbb{R}_+^n}{\operatorname{argmin}} \|R\mathbf{y} - \mathbf{b}\|^2,\end{aligned}$$

with both $f_1(\mathbf{x}) = \|H\mathbf{x}\|_1$ and $f_1(\mathbf{x}) = f_{\text{TV}}(\mathbf{x})$. Below we describe the parameter selection and explain the rationale supporting each choice.

The bilevel algorithms were compared against the Fast Iterative Soft-Thresholding Algorithm (FISTA) [2] for the function

$$\|R\mathbf{x} - \mathbf{b}\|^2 + \iota_{\mathbb{R}_+^n}(\mathbf{x}),$$

where ι is the indicator function (23). We have used the starting image described in the next paragraph. We iterated FISTA with a constant stepsize $\lambda = 3$, for the reasons explained in the paragraph following the next.

Starting image The initial guess \mathbf{x}_0 used in the experiments presented in this subsection was the zero image.

Stepsize sequences The sequence $\{\lambda_k\}$ was set to be

$$\lambda_k = \frac{\lambda}{(k+1)^{0.1}}, \quad (44)$$

where λ was chosen to be just small enough so that the squared residual of the first iterations of the one-level algorithm is decreasing. In the examples, $\lambda = 3$ worked well. This choice for λ was noticed to provide the fastest primary objective function decrease during the iterations when using stepsize sequences of the form (44) above. Sequence $\{\eta_k\}$ was given by

$$\eta_k = \frac{10^6}{(k+1)^{0.1}},$$

which was chosen to attain very high values in order not to negatively influence convergence speed by limiting the values of λ_{i_k} or ξ_k (both of which remained constant throughout the iterations of Algorithm 1, without detected need for decreasing). Finally, the stepsize sequence $\{\mu_k\}$ was given by

$$\mu_k = \frac{\mu}{k+1}.$$

Now, the initial stepsize μ has been selected in order to make the first pair of subiterations to have a specific relative strength. The precise procedure was the following: we first computed $\mathbf{x}_{1/3}$ as usual, following Algorithm 1 and using the already chosen λ_0 . Then, a tentative subiteration $\tilde{\mathbf{x}}_{2/3}$ is computed using a tentative stepsize $\tilde{\mu}_0 = 1$. Finally, the value to be used as starting stepsize is computed by

$$\mu = 10^{-2} \frac{\|\mathbf{x}_0 - \mathbf{x}_{1/3}\|}{\|\mathbf{x}_{1/3} - \tilde{\mathbf{x}}_{2/3}\|}. \quad (45)$$

So, the step given by the first subiteration for the primary problem is about 10^2 times the step given by the subiteration for the secondary optimization problem. This value provided a good compromise between primary problem convergence and secondary function value during the iterations.

Secondary Operators When the problem being solved had $f_1 = f_{\text{Haar}}$, we used FIBA with $\mathcal{O}_{f_1} = \mathcal{N}_{f_{\text{Haar}}}$ according to (38). If the problem had $f_1 = f_{\text{TV}}$, then we have used $\mathcal{O}_{f_1} = \hat{\mathcal{O}}_{f_{\text{TV}}}^{10}$ as defined in (42) with $\hat{\mathcal{O}}_{f_{\text{TV}}}(\lambda, \mathbf{x}) = \mathbf{x} - \lambda \tilde{\nabla} f_{\text{TV}}(\mathbf{x})$ where $\tilde{\nabla} f_{\text{TV}}(\mathbf{x}) \in \partial f_{\text{TV}}(\mathbf{x})$.

Algorithm Convergence Because all entries of the matrix R are nonnegative and because every pixel of the reconstructed image is crossed by several rays during acquisition, the model has the property that X_1 is bounded. Therefore, given the subgradient boundedness of all involved objective functions and the fact that both secondary objective functions f_{Haar} and f_{TV} are bounded from below, Theorem 1 could be applied to show that Algorithm 1 converges in the cases covered in the present subsection, except for two issues. The first problem is that λ_0 may not be smaller than $1/L_0$, but we have observed during algorithm evaluation that the alternative sufficient decrease criterion (22) was satisfied in every iteration, which suffices, instead, for convergence. Furthermore, boundedness of the iterates was also observed, ensuring convergent behavior without forceful truncation.

Numerical Results In the plots that follow, we refer to FIBA when applied to the model with $f_1 = f_{\text{Haar}}$ as FIBA-H and when applied to the model with $f_1 = f_{\text{TV}}$ as FIBA-TV. Figure 3 shows that, as expected, the bilevel approach influences the iterates of the method, thereby resulting in lower secondary objective function value throughout the iterations. Figure 4, on the other hand and also unsurprisingly, shows that the convergence speed for the primary optimization problem is reduced as the secondary optimization step influences the iterations. Figure 5 displays some resulting images, all with the same prescribed primary objective function value, from the algorithms and Figure 6 brings a plot of the profiles across the same line of each image.

We can see the benefits of the enforced smoothing, in particular the good quality of the Total Variation image. The advantage of the bilevel technique here is that the regularization level is chosen by a stopping criterion instead of a parameter in an optimization problem. This latter situation implies that more computational effort would be required since reconstruction for each tentative parameter value would require the equivalent to a large fraction of the computation effort of performing some iterations of the bilevel techniques. We further would like to remark that although heuristic, the accelerating scheme does substantially speed up the method. While we do not show these results in the present paper, when using a non-accelerated technique (i.e, $\eta_k = 0$), the convergence speed is reduced to a rate similar to the ISTA [2] method.

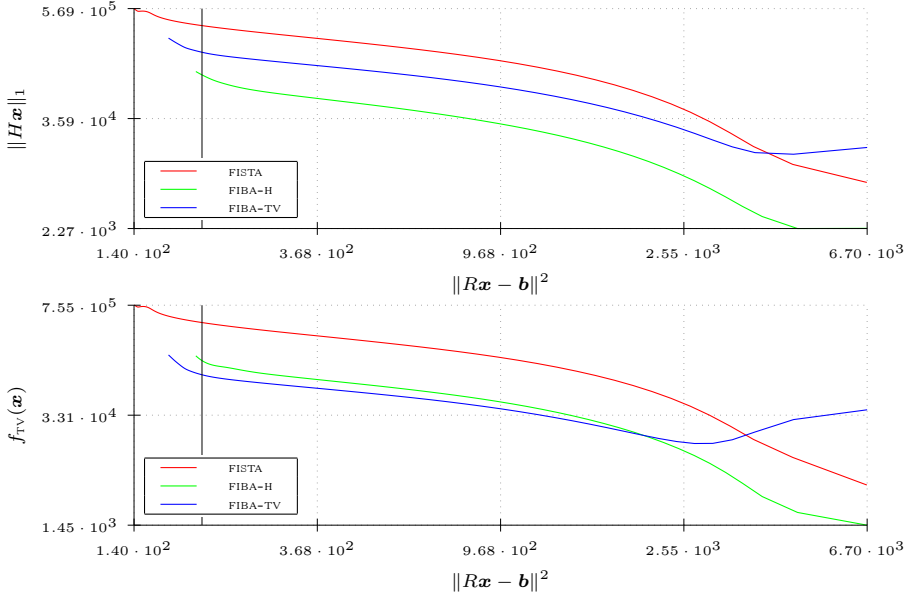


Fig. 3 Logarithmic-scale plots of the trajectories followed by the three studied algorithms over the “phase-plane” described by $f_0 \times f_1$. The solid vertical line depicts the residual value of the images shown at Figure 5. In both graphics, horizontal scale is $\|R\mathbf{x} - \mathbf{b}\|^2$. On top, vertical axis depicts $\|H\mathbf{x}\|_1$. On bottom, curve height is proportional to $f_{TV}(\mathbf{x})$.

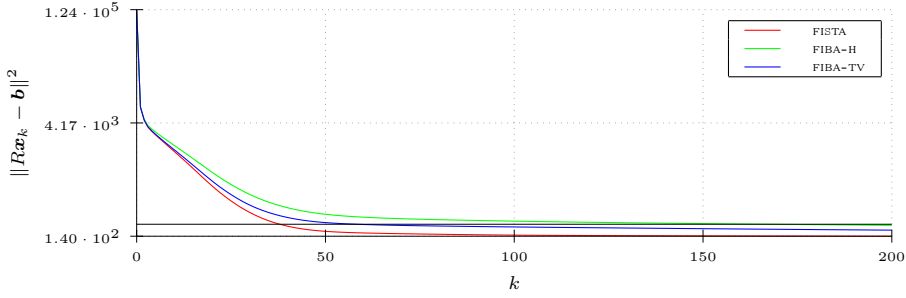


Fig. 4 Evolution of primary objective function over iterations. Solid horizontal line: residual value of the images shown in Figure 5.

4.2 Non-Differentiable Primary Objective Function

Now we consider using Algorithm 2 applied to optimization problem

$$\begin{aligned} \min \quad & f_1(\mathbf{x}) \\ \text{s. t.} \quad & \mathbf{x} \in \operatorname{argmin}_{\mathbf{y} \in \mathbb{R}_+^n} \|R\mathbf{y} - \mathbf{b}\|_1, \end{aligned}$$

with $f_1(\mathbf{x}) = f_{TV}(\mathbf{x})$. We do not experiment using $f_1(\mathbf{x}) = \|H\mathbf{x}\|_1$ here, we leave it to the simulated experiments presented in the next subsection.

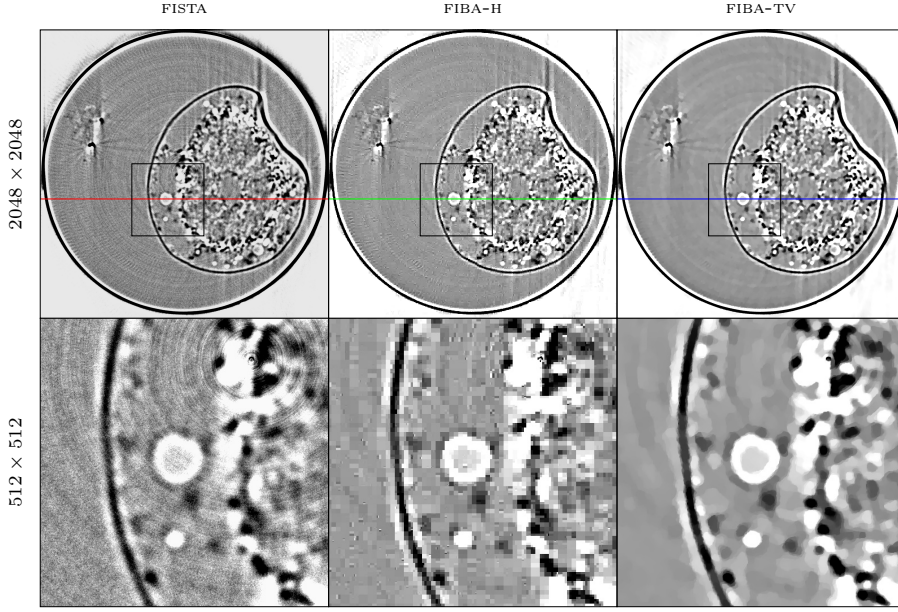


Fig. 5 Reconstructions of fish egg slice from synchrotron radiation transmission data. Top-row: full slice image. Bottom row: details, with location in respective images above shown as a solid black square. The colored solid lines in the top images show the position of the profiles detailed at figure 6.

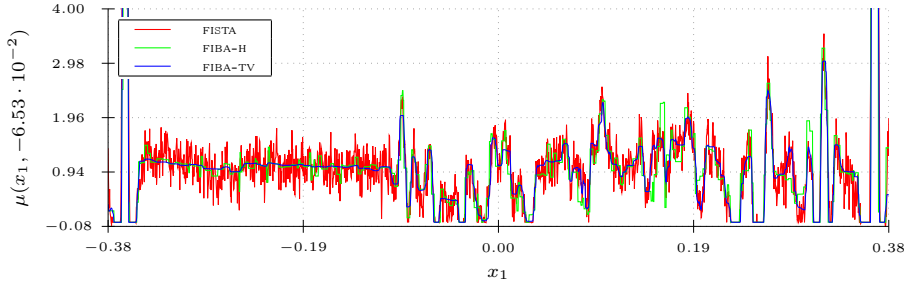


Fig. 6 Profiles through the lines indicated in Figure 5. It is noticeable the noise suppression characteristics of the Total Variation functional, while retaining image detail.

Reconstructions were performed using algorithms where the data was divided in s subsets with $s \in \{1, 2, 4, 8, 16, 32\}$. Each subset was itself composed by data comprising several projection measurements in sequence, i.e., each subset was a vertical stripe of the image² shown in the bottom of Figure 2. At each iteration, the sequence of subset processing was selected by a pseudo-random shuffling. This makes the algorithm non-deterministic, but still covered by the theory because every subset was used once in every iteration and each of the corresponding subdifferentials is uniformly bounded. Furthermore, we have

² Coincidentally, the vertical colored lines delimit the subsets for the case $s = 4$.

observed a consistent behavior among runs and we had not observed a run where sequential data processing led to better convergence than the random ordering.

The various incremental bilevel algorithms were pairwise compared against their pure projected incremental subgradient counterparts, i.e., a variation of Algorithm 2 with $\mu_k \equiv 0$. All of the methods were started with an uniform image as described for the differentiable model. We will denote the bilevel algorithms by IIBA- s , where s is the number of subsets, and the projected incremental method by INC- s .

Starting image The initial guess \mathbf{x}_0 used in the experiments of the present section, for all algorithms tested, was a constant image such that $\sum_{i=1}^m (R\mathbf{x}_0)_i = \sum_{i=1}^m b_i$. It is easy to compute the correct constant value α from $\alpha = \sum_{i=1}^m b_i / \sum_{i=1}^m (R\mathbf{1})_i$ where $\mathbf{1}$ is the vector, of appropriate dimension, with every coordinate equal to 1. This choice makes sure that the Radon consistency condition $\sum_{i=1}^m (R\mathbf{x})_i = \sum_{i=1}^m b_i$ is satisfied for the first iteration, potentially avoiding large oscillations in the first steps of the algorithm.

Stepsize sequences The sequence $\{\lambda_k\}$ for the incremental algorithms with s subsets was set to be

$$\lambda_k = \frac{\lambda}{(k+1)^{\epsilon_s}},$$

where λ was of the form

$$\lambda = \alpha_s \frac{s f_0(\mathbf{x}_0)}{\|\tilde{\nabla} f_0(\mathbf{x}_0)\|^2},$$

and $\tilde{\nabla} f_0(\mathbf{x}_0) \in \partial f(\mathbf{x}_0)$. The pair (α_s, ϵ_s) was selected through a simple search procedure as follows. Let us denote by $\mathbf{x}_{(s, \alpha, \epsilon)}$ the first iteration to be completed past 4 seconds of computation time by the pure projected incremental algorithm with s subsets (that is, by INC- s) and using parameters (α, ϵ) . Then (α_s, ϵ_s) was given by

$$(\alpha_s, \epsilon_s) := \underset{(\alpha, \epsilon) \in \{0.1, 0.2, \dots, 1.0\} \times \{0.5, 0.6, \dots, 0.9\}}{\operatorname{argmin}} f_0(\mathbf{x}_{(s, \alpha, \epsilon)}).$$

The parameters were only optimized for the non-bilevel case, and the same values were used for the corresponding (with relation to the number of subsets) bilevel algorithm. The second step sequence $\{\mu_k\}$ was prescribed as

$$\mu_k = \frac{\mu}{(k+1)^{\epsilon_s+0.1}},$$

with

$$\mu = 10^{-1} \frac{\|\mathbf{x}_0 - \mathbf{x}_{1/3}\|}{\|\mathbf{x}_{1/3} - \tilde{\mathbf{x}}_{2/3}\|},$$

where the rationale is the same than in (45), but with target relative importance between the first subiterations of 10.

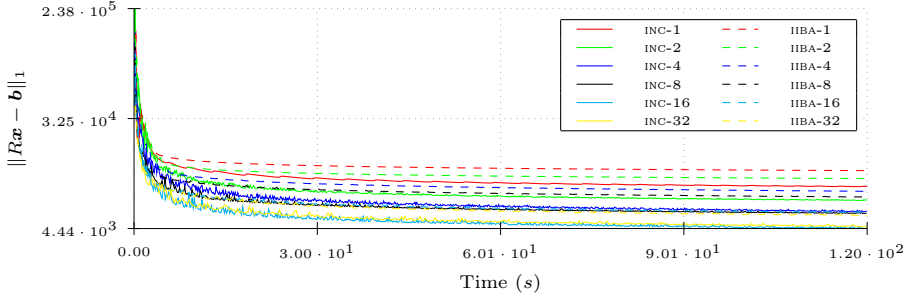


Fig. 7 Convergence of incremental and incremental bilevel algorithms: primary objective function value as a function of computation time.

Secondary Operators In the present experiments we have used $\mathcal{O}_{f_1} = \hat{\mathcal{O}}_{f_{TV}}^5$ as defined in (42) with $\hat{\mathcal{O}}_{f_{TV}}(\lambda, \mathbf{x}) = \mathbf{x} - \lambda \tilde{\nabla} f_{TV}(\mathbf{x})$ where $\tilde{\nabla} f_{TV}(\mathbf{x}) \in \partial f_{TV}(\mathbf{x})$.

Algorithm Convergence Notice that the stepsizes are of the form

$$\lambda_k = \frac{\lambda}{(k+1)^\epsilon} \quad \text{and} \quad \mu_k = \frac{\mu}{(k+1)^{\epsilon+0.1}},$$

where $\epsilon \in [0.5, 0.9]$ and λ and μ are nonnegative. It is routine to check that for this range of ϵ there holds:

$$\sum_{k=0}^{\infty} \lambda_k = \sum_{k=0}^{\infty} \mu_k = \infty, \quad \frac{\mu_k}{\lambda_k} \rightarrow 0 \quad \text{and} \quad \frac{\lambda_k^2}{\mu_k} \rightarrow 0.$$

Also, for the same reasons as in the differentiable primary problem case, the model has the property that X_1 is bounded. Given the subgradient boundedness of all involved objective functions and the fact that secondary objective function f_{TV} is bounded from below, Theorem 2 can be applied to prove algorithm convergence.

Numerical Results By denoting R_i , $i \in \{1, 2, \dots, s\}$, the matrix with the rows corresponding to the i -th subset, we notice that the computationally demanding parts of the algorithm are products of the form

$$R_i \mathbf{x} \quad \text{and} \quad R_i^T \mathbf{y},$$

because the partial subgradients are given by

$$R_i^T \mathbf{sign}(R_i \mathbf{x} - \mathbf{b}).$$

We were not able to make the NFFT library as efficient for such partial matrix-vector products, which imposed a large overhead in the partial iterations. We have instead used a ray-tracing algorithm [20, 33] implemented to run in GPUs (Graphics Processing Units) under single precision floating point arithmetics.

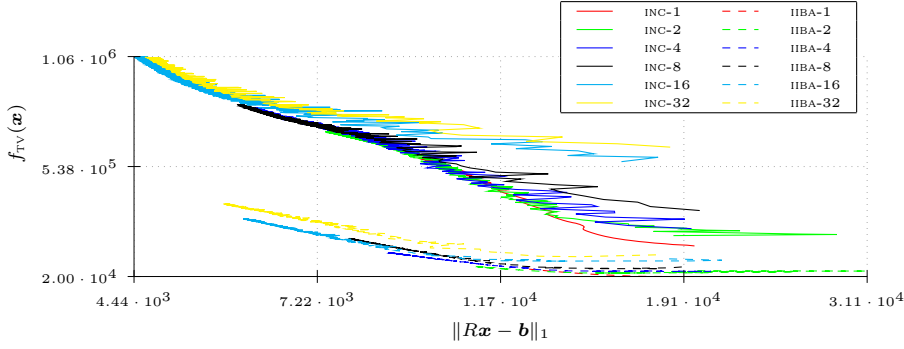


Fig. 8 Total Variation versus residual norm-1. Notice that the bilevel algorithms present significantly better f_{TV} values than those for the original model. On the other hand, it is seen here and in Figure 7 that, for example, IIBA-32 is competitive, in terms of f_0 reduction, with INC-4 while still maintaining a considerable better f_1 value than the latter, for the same f_0 value.

A special scheme was used so that access to slow GPU memory is minimized by performing the computations in sub-images loaded to/from the GPU's shared memory (a 64kb fast L1-cache-like memory) in coalesced reads/writes, and summing up the partial results. In this setting, the sequential computation of the s different partial subgradients takes longer than the computation of the subgradient itself, because there are more memory copy to/from shared memory. Yet another per-iteration overhead of the incremental methods are the multiple subiteration updates. Therefore one iteration of the incremental method with s subsets still takes slightly longer than with $s - 1$ subsets.

Even with mandatory overheads, a careful implementation was able to make the iteration-wise speed up provided by the incremental approach advantageous time-wise, as can be seen in Figure 7. An important feature in this plot is that this speed up is retained by the bilevel algorithms in a similar fashion to the non-bilevel incremental method. That is, if we take into account that considering the secondary objective function in the optimization does, expectedly, slow down the non-incremental method from the viewpoint of primary objective function decrease in comparison to the corresponding non-bilevel algorithm, it can be seen that incremental bilevel techniques too present a speed up in this convergence rate as the number of subsets grow. Computations were performed on a GTX 480 GPU and timing figures were obtained considering iterate updates only, disregarding both data input/output and objective value computation.

A particular point in the experimental results can be seen in Figure 8. In the application of tomographic reconstruction, incrementalism seems to induce more roughness and we therefore see that INC- s achieves lower Total Variation for the same value of 1-norm of the residual than INC- $2s$. Looking at the IIBA- s curves in the same plot, we notice that the choice of secondary objective function, which as we have seen is conflicting with the incrementality idea in the one level case, also plays a similar role in the bilevel case. Consequently, at

least for our algorithmic parameters selection, there is a incrementality level (or equivalently, primary objective function decrease speed) versus secondary objective function decrease trade-off. Even so, IIBA-32 provides substantially lower Total Variation for a given 1-norm of the residual compared to INC- s for every s while still achieving faster experimental primary objective function decrease rate than INC- s with $s \leq 4$.

For other bilevel models, at first glance, there seems not to be any reason for an increase in incrementality to lead to worse secondary to primary objective function ratios. However, such antagonism between fast algorithms (with relation to data adherence) and desirable solution properties may appear naturally in models for ill-posed inverse problems like the one we consider here, because in this case overly fit solutions to noisy data are unstable and the secondary objective function is usually an attempt at instability prevention.

4.3 Simulated Data

The present set of experiments intends to highlight the practical differences and advantages of the bilevel approach

$$\begin{aligned} \min \quad & \|H\mathbf{x}\|_1 \\ \text{s. t.} \quad & \mathbf{x} \in \underset{\mathbf{y} \in \mathbb{R}^n}{\operatorname{argmin}} \|R\mathbf{y} - \mathbf{b}\|^2 \end{aligned} \quad (46)$$

over regularized techniques of the form

$$\min \quad \frac{1}{2} \|R\mathbf{y} - \mathbf{b}\|^2 + \gamma \|H\mathbf{x}\|_1. \quad (47)$$

In order to be able to quantify reconstructed image quality, we use simulated data in these experiments. The ideal image will be denoted by \mathbf{x}^\dagger and was a 512×512 pixels discretized and scaled version of the Shepp-Logan phantom that can be seen at the left of Figure 1. The scaling was such that the relative error after Poisson data simulation was around 10%. Tomographic data was computed at 64 angular samples evenly spaced in $[0, \pi)$, each by its turn sampled in 512 points in $[-1, 1]$.

We have used FIBA for solving (46) with just the same parameters (including the secondary operator $\mathcal{O}_{f_1} = \mathcal{N}_{f_{\text{Harr}}}$) of those used in Subsection 2.4 except that convergence required $\lambda = 2$ and a reasonable starting point for the secondary stepsize sequence was $\mu = 10^2$. Problem (47) was solved by the FISTA algorithm with a constant stepsize $\lambda = 2$. We have tried $\gamma \in \{10^2, 10, 1.5, 1, 0\}$. We have run both algorithms for 400 iterations in this experiment and, as in Subsection 2.4, the algorithms were started with the zero image.

Figure 10 shows how the image quality, as measured by the relative error $\|\mathbf{x}_k - \mathbf{x}^\dagger\|/\|\mathbf{x}^\dagger\|$, evolves over the iterations for the tested methods. In Figure 9 we see the best image obtained by each of the methods throughout the iterations. We notice that the BILEVEL image is competitive with the FISTA for a certain range of values of γ , whereas if γ is not within the reasonable

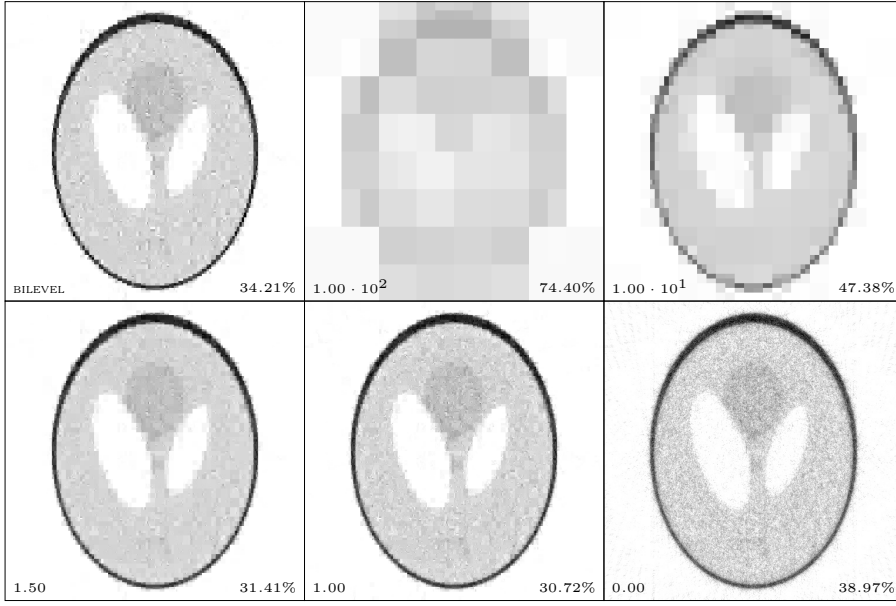


Fig. 9 Reconstructions of the Shepp-Logan phantom from simulated noisy data. Top left: best image obtained during execution of FIBA. Other images are the best obtained during executions of FISTA, in these images the bottom-left label is the value of the regularization parameter. The bottom-right label is the relative image error. Relative data error was 10.19%.

range, reconstruction by solving (47) quickly degrades. In fact, had we used larger values of the starting secondary stepsize good images would still be obtained, maybe requiring more iterations, while larger than ideal values for γ in the non-bilevel approach produce a wholly unusable sequence of iterates. On the other hand, smallish secondary stepsize sequences in the bilevel method have practically the same effect than using a small γ in the non-bilevel approach, except that the notion of “small” includes a wider range of values in the bilevel approach because of the diminishing nature of the sequence $\{\mu_k\}$. Therefore, if we consider the starting secondary stepsize μ a parameter of the bilevel technique, it is considerably easier to choose than the parameter γ of the traditional regularization approach. On the other hand, if a good and efficient procedure for selecting γ is available and the regularization function is appropriate, solving (47) has the potential of delivering better reconstructions.

5 Conclusions

The present paper introduced an abstract class of explicit numerical methods for instances of bilevel non-differentiable convex optimization problems and proposed two first-order concrete representatives of these new algorithms with reduced per iteration computational cost. The proposed methods have computationally simple iterations. The reported numerical experimentation

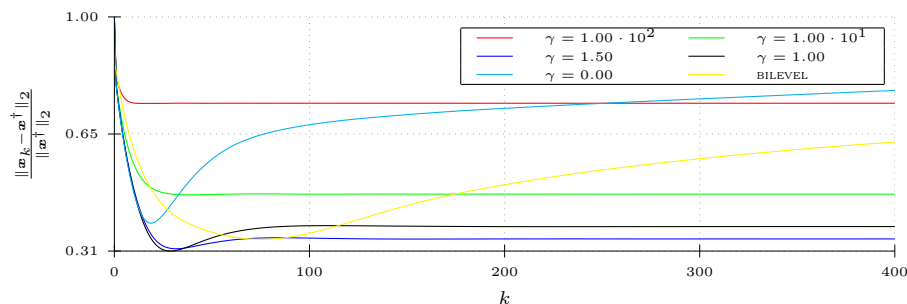


Fig. 10 Image quality evolution as iterations proceeds in simulated experiment.

showed that, when used in tomographic reconstruction problems where the problem size is huge, the low computational complexity of the iterations and the advanced modeling allow to the technique to generate high quality images at a moderate computational cost from sparsely sampled and noisy data. Algorithmic flexibility was highlighted by two conceptually distinct implementations: on one side, an implementation in the high-level Python language, which is free and portable to most computing architectures and environments. On the other hand, the simplicity of the method was also suitable for a low-level hardware-specific implementation fully running on a GPU with no external software library dependency.

Acknowledgements

We would like to thank the LNLS for providing the beam time for the tomographic acquisition, obtained under proposal number 17338. We are also grateful to Prof. Marcelo dos Anjos (Federal University of Amazonas) for kindly providing the fish egg samples used in the presented experimentation and Dr. Eduardo X. Miqueles for invaluable help in data acquisition and for the discussions about tomographic reconstruction. We are also indebted to the anonymous referees who gave numerous suggestions that lead to the improvement of the original manuscript.

References

1. AMIR BECK AND SHOHAN SABACH. A first order method for finding minimal norm-like solutions of convex optimization problems. *Mathematical Programming*, 147(1):25–46, 2014. doi:10.1007/s10107-013-0708-2.
2. AMIR BECK AND MARC TEBoulLE. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi:10.1137/080716542.
3. DIMITRI P. BERTSEKAS. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011. doi:10.1007/s10107-011-0472-0.

4. DIMITRI P. BERTSEKAS. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997. doi:10.1137/S1052623495287022.
5. DIMITRI P. BERTSEKAS AND JOHN N. TSITSIKLIS. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. doi:10.1137/S1052623497331063.
6. DORON BLATT, ALFRED O. HERO AND HILLEL GAUCHMAN. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007. doi:10.1137/040615961.
7. KRISTIAN BREDIES AND MARIYA ZHARIY. A discrepancy-based parameter adaptation and stopping rule for minimization algorithms aiming at Tikhonov-type regularization. *Inverse Problems*, 29(2):025008, 2013. doi:10.1088/0266-5611/29/2/025008.
8. ALEXANDRE CABOT. Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchical minimization. *SIAM Journal on Optimization*, 15(2):555–572, 2005. doi:10.1137/S105262340343467X.
9. EMMANUEL J. CANDÈS, JUSTIN K. ROMBERG AND TERENCE TAO. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. doi:10.1109/TIT.2005.862083.
10. EMMANUEL J. CANDÈS, JUSTIN K. ROMBERG AND TERENCE TAO. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006. doi:10.1002/cpa.20124.
11. JON F. CLAERBOUT AND FRANCIS MUIR. Robust modelling with erratic data. *Geophysics*, 38(5):826–844, 1973. doi:10.1190/1.1440378.
12. RAFAEL CORREA AND CLAUDE LEMARÉCHAL. Convergence of some algorithms for convex minimization. *Mathematical Programming*, 62:261–275, 1993. doi:10.1007/BF01585170.
13. ÁLVARO RODOLFO DE PIERRO AND MICHEL EDUARDO BELEZA YAMAGISHI. Fast EM-like methods for maximum “a posteriori” estimates in emission tomography. *IEEE Transactions on Medical Imaging*, 20(4):280–288, 2001. doi:10.1109/42.921477.
14. D. L. DONOHO AND B. F. LOGAN. Signal recovery and the large sieve. *SIAM Journal on Applied Mathematics*, 52(2):577–591, 1992. doi:10.1137/0152031.
15. DAVID L. DONOHO. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi:10.1109/TIT.2006.871582.
16. DAVID L. DONOHO. For most large underdetermined systems of linear equations, the minimal ℓ_1 -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006. doi:10.1002/cpa.20132.
17. DAVID L. DONOHO. For most large underdetermined systems of equations, the minimal ℓ_1 -norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006. doi:10.1002/cpa.20131.

18. KARSTEN FOURMONT. Non-equispaced fast Fourier transforms with applications to tomography. *Journal of Fourier Analysis and Applications*, 9(5): 431–450, 2003. doi:10.1007/s00041-003-0021-1.
19. EDGAR GARDUÑO AND GABOR T. HERMAN. Superiorization of the ML-EM algorithm. *IEEE Transactions on Nuclear Science*, 61(1):162–172, 2014. ISSN 00189499. doi:10.1109/TNS.2013.2283529.
20. GUOPING HAN, ZHENGRONG LIANG AND JIANGSHENG YOU. A fast ray-tracing technique for TCT and ECT studies. *Conference Records of the 1999 IEEE Nuclear Science Symposium*, (3):1515–1518, 1999. doi:10.1109/NSSMIC.1999.842846.
21. ELIAS S. HELOU, YAIR CENSOR, TAI-BEEN CHEN, I-LIANG CHERN, ÁLVARO R. DE PIERRO, MING JIANG AND HENRY H.-S. LU. String-averaging expectation-maximization for maximum likelihood estimation in emission tomography. *Inverse Problems*, 30(5):055003, 2014. doi:10.1088/0266-5611/30/5/055003.
22. ELIAS SALOMÃO HELOU NETO AND ÁLVARO RODOLFO DE PIERRO. On perturbed steepest descent methods with inexact line search for bilevel convex optimization. *Optimization*, 60(8-9):991–1008, 2011. doi:10.1080/02331934.2010.536231.
23. GABOR T. HERMAN. *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*. Academic Press, 1980.
24. PETER J. HUBER. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. URL <http://projecteuclid.org/euclid.aoms/1177703732>.
25. AVINASH C. KAK AND MALCOLM SLANEY. *Principles of Computerized Tomographic Imaging*. IEEE press, 1988.
26. JAMES KEINER, STEFAN KUNIS AND DANIEL POTTS. Using NFFT3 – a software library for various nonequispaced fast Fourier transforms. *ACM Transactions on Mathematical Software*, 2009. doi:10.1145/1555386.1555388.
27. EDUARDO X. MIQUELES, JEAN RINKEL, FRANK O'DOWD AND JUAN S. V. BERMÚDEZ. Generalized Titarenko's algorithm for ring artefacts reduction. *Journal of Synchrotron Radiation*, 21:1333–1346, 2014. doi:10.1107/S1600577514016919.
28. FRANK NATTERER. *The Mathematics of Computerized Tomography*. Wiley, 1986.
29. ANGELIA NEDIĆ AND DIMITRI P. BERTSEKAS. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001. doi:10.1137/S1052623499362111.
30. R. TYRRELL ROCKAFELLAR. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. doi:10.1137/0314056.
31. LEONID I. RUDIN, STANLEY OSHER AND EMAD FATEMI. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992. doi:10.1016/0167-2789(92)90242-F.
32. FADIL SANTOSA AND WILLIAM W. SYMES. Linear inversion of band-

- limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986. doi:10.1137/0907087.
33. ROBERT L. SIDDON. Fast calculation of the exact radiological path for a three-dimensional CT array. *Medical Physics*, 12(2):252–255, 1985. doi:10.1118/1.595715.
34. MIKHAIL SOLODOV. An explicit descent method for bilevel convex optimization. *Journal of Convex Analysis*, 14(2):227–237, 2007. URL <http://www.heldermann.de/JCA/JCA14/JCA142/jca14016.htm>.
35. MIKHAIL SOLODOV. A bundle method for a class of bilevel nonsmooth convex minimization problems. *SIAM Journal on Optimization*, 18(1):242–259, 2008. doi:10.1137/050647566.
36. MIKHAIL V. SOLODOV. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998. doi:10.1023/A:1018366000512.
37. MIKHAIL V. SOLODOV AND S. K. ZAVRIEV. Error stability properties of generalized gradient-type algorithms. *Journal of Optimization Theory and Applications*, 98(3):663–680, 1998. doi:10.1023/A:1022680114518.
38. HOWARD L. TAYLOR, STEPHEN C. BANKS AND JOHN F. MCCOY. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979. doi:10.1190/1.1440921.